Proceedings of the 2nd International Workshop on Computational Social Choice (COMSOC-2008)

Ulle Endriss & Paul W. Goldberg (eds.)

Department of Computer Science
University of Liverpool

Programme Committee

Felix Brandt	University of Munich, Germany
Vincent Conitzer	Duke University, United States
Edith Elkind	University of Southampton, United Kingdom
Ulle Endriss (*co-chair*)	University of Amsterdam, The Netherlands
Paul Goldberg (*co-chair*)	University of Liverpool, United Kingdom
Lane Hemaspaandra	University of Rochester, United States
Sébastien Konieczny	CNRS & Université d'Artois, France
Jérôme Lang	CNRS & Université Paul Sabatier, France
Christian List	London School of Economics, United Kingdom
Vangelis Markakis	CWI, Amsterdam, The Netherlands
Vahab Mirrokni	Google Research, NYC, United States
Gabriella Pigozzi	University of Luxembourg, Luxembourg
Francesca Rossi	University of Padova, Italy
Remzi Sanver	Istanbul Bilgi University, Turkey
Arkadii Slinko	University of Auckland, New Zealand
Michael Trick	Carnegie Mellon University, United States
Berthold Vöcking	RWTH Aachen University, Germany
William S. Zwicker	Union College, United States

Additional Reviewers

Stéphane Airiau, Arash Asadpour, Yoram Bachrach, Elise Bonzon, Yann Chevaleyre, Bruno Escoffier, Piotr Faliszewski, Felix Fischer, Renato Gomes, Paul Harrenstein, Christopher Homan, Rob LeGrand, Hamid Mahini, Pierre Marquis, Nicolas Maudet, Eric Pacuit, Ramón Pino Pérez, Ariel Procaccia, Clemens Puppe, Talal Rahwan, Heiko Roeglin, Tamas Sarlos, François Schwarzentruber, Joel Uckelman, K. Brent Venable, Frank Wolter, and Mike Wooldridge

Local Organising Committee

Wiebe van der Hoek	University of Liverpool, United Kingdom
Peter McBurney	University of Liverpool, United Kingdom
Mike Wooldridge	University of Liverpool, United Kingdom

Sponsor

UK Engineering and Physical Sciences Research Council (EPSRC) —
Market-Based Control of Complex Computational Systems (MBC) Project (GR/T10657/01)

Workshop Website

`http://www.csc.liv.ac.uk/~pwg/COMSOC-2008/`

Preface

Computational social choice is an interdisciplinary field of study bringing together ideas from social choice theory and computer science. It is concerned with the application of techniques developed in computer science, such as complexity analysis or algorithm design, to the study of social choice mechanisms, such as voting procedures or fair division algorithms. It also seeks to import concepts from social choice theory into computing, for instance for the analysis of computer networks or multiagent systems.

These are the proceedings of COMSOC-2008, the 2nd International Workshop on Computational Social Choice, hosted by the Department of Computer Science at the University of Liverpool on 3-5 September 2008. The first COMSOC workshop has been held in 2006 in Amsterdam. The aim of the workshop series is to bring together different communities: computer scientists interested in computational issues in social choice; people working in artificial intelligence and multiagent systems who are using ideas from social choice to organise societies of artificial software agents; logicians interested in the logic-based specification and analysis of social procedures; and last but not least people coming from social choice theory itself. COMSOC is intended to continue as a regular biannual event.

While COMSOC has started as an initiative of people with a background in computer science (and logic and AI), we hope to continue to strengthen ties with the social choice theory community. We are therefore particularly pleased about the significant level of participation of prominent members of this community in the workshop, be it as Programme Committee members, reviewers, invited speakers, or contributing authors.

We received 55 submissions (not including a handful of off-topic submissions that were deleted at the outset). This represents a 15% increase over 2006. Each submission has been reviewed by three members of the Programme Committee, supported by a large number of additional reviewers. We eventually accepted 36 papers out of the 55 submissions for presentation at the workshop.

This volume collects revised versions of these accepted papers, taking the comments of our reviewers into account. It also includes abstracts of the talks by our five invited speakers: Salvador Barberà (Barcelona), Rohit Parikh (CUNY), Tuomas Sandholm (Carnegie Mellon), Moshe Tennenholtz (Technion), and William Thomson (Rochester). As for the first edition of the workshop, the Call for Papers explicitly solicited submissions of both original papers and of papers describing recently published work, so some of the papers have recently appeared also in other publication venues. The reason for this policy has been to try to accommodate the varying publishing traditions of different disciplines and to ensure that COMSOC-2008 would attract a representative sample of the best work in the field. The copyright for the articles in this volume lies with the individual authors.

The programme covers a wide range of topics, ranging from complexity questions in election manipulation, over ranking systems, coalitional voting games and mechanism design, to judgment aggregation. Important topics that have been missing from the 2006 programme and that are included in the present volume are the analysis of tournaments, matching theory, and belief merging. On the other hand, not all of the topics covered in 2006 are again represented in this year's programme. To name just one example, COMSOC-2006 included two sessions on fair division and resource allocation problems. Unfortunately, this year we do not have contributed papers in this area (but the topic still does get some coverage, thanks to the invited talk by William Thomson). We hope that all of these subfields of computational social choice will continue to flourish over the coming years and be well represented at future editions of COMSOC.

We would like to thank all authors for submitting their papers, the invited speakers, presenters and other workshop participants for attending, and the members of the Programme Committee and all additional reviewers for their invaluable help in putting together an exciting scientific programme. Special thanks are also due to Christian List (LSE) and to Jörg Rothe (Düsseldorf) for having agreed to deliver tutorials on the day before the workshop.

Finally, we are grateful for the sponsorship received from the UK Engineering and Physical Sciences Research Council (EPSRC), through the Market-Based Control of Complex Computational Systems (MBC) project (GR/T10657/01).

Amsterdam & Liverpool U.E. & P.W.G.
August 2008

Invited Talks

Contributed Papers

Individual and Group Strategy Proofness of Voting Rules: The Case for Restricted Domains

Salvador Barberà

All nontrivial voting rules are manipulable at some preference profiles when defined on a universal domain. This manipulation is possible even by single individuals, and "a fortiori", by coalitions of voters. This negative result may be reversed if the functions are defined in appropriately restricted domains of preferences, and only preferences in that same domain can be used to manipulate.

In the first part of the talk I will exemplify the type of results regarding non-manipulability by individuals (or strategy-proofness) that one can obtain under domain restrictions. I will pay special attention to the domain of separable preferences on grids. I will present a characterization of all strategy-proof rules in that context, consider the complications that arise in the presence of feasibility constraints and discuss the relevance of that paradigmatic case. I will also briefly touch upon the case of voting rules whose outcomes are lotteries.

In the second part of the talk I will consider the added difficulties that arise if one wants to guarantee that a rule is non-manipulable by coalitions of voters, in addition to being individually strategy-proof. I will also consider intermediate cases, where only "large enough" groups may manipulate. I will also show that, in spite of these difficulties, some domains of preferences have the nice property that all strategy-proof rules that one can define on them are also necessarily group strategy-proof (or at least strategy-proof in front of coalitions that are not "too large"). I will provide a characterization of those nice domains of preferences for which both types of requirements become equivalent.

Salvador Barberà
Departament d'Economia i d'Història Econòmica and CODE
Universitat Autònoma de Barcelona
08193 Bellaterra (Barcelona), Spain
Email: salvador.barbera@uab.es

Talk, Cheap Talk, and States of Knowledge

Rohit Parikh

Applications of Epistemic Logic have by now become a major industry and an area once dominated by philosophers has now attracted large followings among both AI people and economists. We will discuss some of our own work in this area, including applications to various social issues, like consensus, common knowledge, elections, the sorts of things which candidates running for office are apt to say, and why.

Very important issues in conversation and in the working of other interactions are the ways in which states of knowledge and belief change when things happen or when someone says something. There is a great deal of material on this topic where issues like Kripke structure tranformation [3], and history based models [10] enter. In sophisticated applications, Gricean implicature [5], or cheap talk [2, 11] may also enter.

Gricean implicature assumes a co-operative stance, whereas cheap talk is a notion which also makes sense when the interests of the speaker and listener are only partially aligned. Game theoretic considerations become relevant.

States of knowledge and changes in them have social and economic consequences, and there have been developments starting with Aumann's seminal paper [1], followed by work by [4, 9] and others. Milgrom and Stokey's no trade theorem [6] is also an important consequence.

We will give an overview of representations of states of knowledge, of changes in them, and the social consequences.

References

[1] R. Aumann, Agreeing to Disagree, *Annals of Statistics*, **4** (1976), 1236-1239.

[2] Crawford, V. and J. Sobel (1982): Strategic Information Transmission, *Econometrica*, **50**, 1431–1452.

[3] Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi *Dynamic Epistemic Logic* (Synthese Library) (2007)

[4] J. Geanakoplos and H. Polemarchakis, We Can't Disagree Forever, *J. Economic Theory*, **28** (1982), 192-200.

[5] Paul Grice, *Studies in the Way of Words*, Harvard U. Press (1989).

[6] Paul Milgrom and Nancy Stokey, Information, Trade and Common Knowledge. *Journal of Economic Theory* **26** (1982) 17–27

[7] R. Parikh, Sentences, Propositions and Logical Omniscience, to appear in *The Review of Symbolic Logic.*

[8] R. Parikh Knowledge and Structure in Social Algorithms, presented at the 3rd International Game Theory Conference, July 2008.

[9] R. Parikh and P. Krasucki, Communication, Consensus and Knowledge, *J. Economic Theory* **52** (1990) pp. 178-189.

[10] R. Parikh and R. Ramanujam, A Knowledge based Semantics of Messages, *J. Logic, Language and Information* **12** 2003, 453-467

[11] Robert Stalnaker Saying and Meaning, Cheap Talk and Credibility in *Game Theory and Pragmatics* Editors: Anton Benz, Gerhard Jager and Robert van Rooij Palgrave Macmillan, (2005) pp. 83–100

Rohit Parikh
Department of Computer Science
CUNY Graduate Center
New York, NY 10016-4309, USA
Email: rparikh@gc.cuny.edu

Expressiveness in Mechanisms and its Relation to Efficiency: Our Experience from $40 Billion of Combinatorial Multi-attribute Auctions, and Recent Theory

Tuomas Sandholm

Abstract

A recent trend (especially in electronic commerce) is higher levels of expressiveness in the mechanisms that mediate interactions such as auctions, exchanges, catalog offers, voting systems, matching of peers, and so on. Participants can express their preferences in drastically greater detail than ever before. In many cases this trend is fueled by modern algorithms for winner determination that can handle the richer inputs. But is more expressiveness always a good thing? What forms of expressiveness should be offered? In this talk I will first report on our experience from over $40 billion of combinatorial multi-attribute sourcing auctions. Then, I will present recent theory that ties the expressiveness of a mechanism to an upper bound on efficiency in a domain-independent way in private-information settings. Time permitting, I will also discuss theory and experiments on applying expressiveness to ad auctions, such as sponsored search and real-time banner ad auctions with temporal span and complex preferences.

1 Introduction

By carefully crafting mechanisms it is possible to design better auctions, exchanges, catalog offers, voting systems, and so on. A recent trend in the world—especially in electronic commerce—is a demand for higher levels of expressiveness in the mechanisms that mediate interactions such as the allocation of resources, matching of peers, or elicitation of privacy and security preferences.

The most famous expressive mechanism is a *combinatorial auction (CA)*, which allows participants to express valuations over *packages* of items. CAs have the recognized benefit of removing the exposure problems that bidders face when they have preferences over packages but in traditional auctions are allowed to submit bids on individual items only. CAs also have other acknowledged benefits.

Expressiveness also plays a key role in *multi-attribute* settings where the participants can express preferences over vectors of attributes of the item—or, more generally, of the outcome.

The trend toward expressiveness is also reflected in the richness of preference expression offered by businesses as diverse as matchmaking sites, sites like Amazon and Netflix, and services like Google's AdSense. In Web 2.0 parlance, this demand for increasingly diverse offerings is called the Long Tail [1].

2 Our real-world experiences with expressive mechanisms in sourcing

In the first part of the talk, I will share some of my experiences from using expressiveness in practice. I started building winner determination algorithms for combinatorial auctions

in 1997, and founded a company, CombineNet, Inc., in 2000 to field expressive mechanisms. Since then we have fielded over 500 expressive auctions. These auctions have been in the area of strategic sourcing, that is, the process by which large companies buy materials, products, services, and transportation from their suppliers, striking long-term contracts based on each auction.

Our auction designs, which we now call *expressive commerce*, hybridize and generalize both combinatorial and multi-attribute auctions [7, 9]. Expressive commerce combines the advantages of highly expressive human negotiation with the advantages of electronic reverse auctions. The idea is that supply and demand are expressed in drastically greater detail than in traditional electronic auctions, and are algorithmically cleared. This creates an efficiency improvement in the allocation (a win-win between the buyer and the sellers), but the market clearing problem is a highly complex combinatorial optimization problem. We developed the fastest custom tree search algorithms for solving it. We have hosted over $40 billion of sourcing using the technology, and created over $5 billion of hard-dollar savings plus numerous harder-to-quantify benefits. The suppliers also benefited by being able to express production efficiencies and creativity, and through exposure problem removal.

We found that the traditional form of expressive bidding in CAs, package bidding (possibly with different forms of exclusivity constraints between bids), is a much too impoverished a bidding language to be usable in practice. In contrast, we found that there are a host of more compact and natural expressiveness constructs, and they are all used in concert in our auctions. These include various flexible forms of package bids, rich forms of conditional discount offers, various forms of discount schedules, side constraints, expressions of cost drivers, and multiattribute bidding [7].

In our events the bid taker can also express various forms of preferences and constraints. By conducting what-if analysis by changing these, the bid taker can form a quantitative understanding of the tradeoffs available in the supply chain, such as cost versus multiple measures of practical implementability of the allocation, cost versus multiple measures of quality of the allocation, and cost versus multiple measures of long-term risk entailed by the allocation [7].

Furthermore, by allowing expressive offers over different combinations of the items to be sourced, the winner determination, as a side effect, ends up redesigning the supply chain. For example, in a sourcing event where Procter & Gamble sourced in-store displays using our hosting service and technology, we sourced items from different levels of the supply chain in one event: buying colorants and cardboard of different types, buying the service of printing, buying the transportation, buying the installation service, etc. [8]. Some suppliers made offers for some of those individual items while others offered complete ready-made displays (which are, in effect, packages of the lower-level items), and some bid for partial combinations. The market clearing determined the lowest-cost (adjusted for the Procter & Gamble's constraints and preferences) solution and thus, in effect, configured the supply chain multiple levels upstream.

An additional interesting aspect of bidding with cost drivers and alternates (e.g., using attributes) is that the winner determination algorithm not only decides who wins, but also ends up optimizing the configuration (setting of attributes) for each item, and the process by which each item is made.

3 Theory

Intuitively, one would think that increases in expressiveness would lead to more efficient mechanisms. That is also what the CombineNet experiences suggest. However, until now we have lacked a general theory that ties expressiveness and efficiency.

We developed a theory that ties the expressiveness of mechanisms to their efficiency in a domain-independent manner [3]. We introduce two new expressiveness measures, 1) *maximum impact dimension*, which captures the number of ways that an agent can impact the outcome, and 2) *shatterable outcome dimension*, which is based on the concept of *shattering* from computational learning theory. We derive an upper bound on the expected efficiency of any mechanism under its most efficient Nash equilibrium. Remarkably, it depends only on the mechanism's expressiveness. We prove that the bound increases strictly as we allow more expressiveness. We also show that in some cases a small increase in expressiveness yields an arbitrarily large increase in the bound.

Finally, we study *channel-based* mechanisms. The restriction is that these mechanisms take expressions of value through channels from agents to outcomes, and select the outcome with the largest sum. (Channel-based mechanisms subsume most combinatorial and multi-attribute auctions, the Vickrey-Clarke-Groves mechanism, etc.) In this class, a natural measure of expressiveness is the number of channels allowed (this generalizes the k-wise dependence measure of expressiveness used in the combinatorial auction literature). We show that our domain-independent measures of expressiveness appropriately relate to the natural measure of expressiveness of channel-based mechanisms: the number of channels allowed. Using this bridge, our general results yield interesting implications. For example, any (channel-based) multi-item auction that does not allow rich combinatorial bids can be arbitrarily inefficient—unless agents have no private information.

4 Applications to ad auctions and exchanges

Advertisement auctions and exchanges are relatively new forms of buying and selling ad space. They are an opportune next area of application for expressive mechanisms.

4.1 The case of an isolated sponsored search auction

Sponsored search auctions (the dispatch of typically textual ads in response to keyword-based web searches) account for tens of billions of dollars in revenue annually (e.g., to Google, Yahoo!, and Microsoft) and are some of the fastest growing mechanisms on the Internet. However, the most frequent variant of these mechanisms does not allow bidders to offer a separate bid for each ad position, and is thus inexpressive on a fundamental level. Here we attempt to characterize the cost of this inexpressiveness [2]. We adapt the theoretical framework discussed in the previous section to show that the commonly used *generalized second price (GSP)* mechanism is arbitrarily inefficient for some distributions over agent preferences. We then describe a search technique that computes an upper bound on the expected efficiency of the GSP mechanism for any given distribution over agent preferences. We report the results of running our search technique on synthetic preference distributions. Our results demonstrate that the cost of inexpressiveness is most severe when agents have diverse preferences (such as having both brand advertisers and value advertisers in the auction) and relatively low profit margins. Our results also show that designating one or more positions as "premium" and soliciting an extra bid for these positions eliminates almost all of the inefficiency.

4.2 Highly expressive real-time ad auctions that span time

The prevalence and variety of online advertising in recent years has led to the development of an array of services for both advertisers and purveyors of media. Because matching an advertiser's needs (demand) with a content provider's properties (e.g., locations on displayed

web pages) is a complex enterprise, often automated matching is used to match ad channels[1] with advertisers. One famous example is sponsored search. Internet auctions of traditional advertising (TV, radio, print) are also emerging (e.g., via companies like *Google* and *Spot Runner*). Auctions and exchanges for banner ads have also been established—e.g., *Right Media* (now part of *Yahoo!*) and *DoubleClick* (now part of *Google*)—although many banner ad bulk contracts are still manually negotiated.

There has been considerable research on developing auction mechanisms for allocating ad channels, with a focus on issues like auction design, charging schemes (e.g., per impression or per *click-through (CT)*), bidder strategies, and so on. However, attention has focused almost exclusively on improving single-period expressiveness, still with per-impression or per-CT prices. As has been well-documented in other auction domains like sourcing, requiring bidders and bid takers to shoehorn their preferences into the impoverished language of per-item bids is usually undesirably restrictive.

Here we explore the use of *expressive bidding* for online banner ad auctions. For ease of presentation, we discuss banner ads, but the general principles and specific techniques we propose can be applied to other forms of online advertising (electronic auctions of TV and radio ads, sponsored search, etc.) as well.

In many domains, the value of a *set* of ads may not be an additive function of value of its individual elements. For instance, in an advertising campaign, *campaign-level* expressiveness is important. Advertisers may value particular *sequences* of ads, rather than individual ads per se. Efficiency (and revenue) maximization in such an environment demand that we allow bidders to express bids (propose *contracts*) on complex allocations, and that bid takers optimize over *sequences* of allocations to best match bidder preferences, in a way that cannot be accommodated using per-item bidding.

The key technical challenge for expressive ad auctions is optimization: determining the optimal allocation of ad channels to very large numbers of complex bids in real-time. This is further complicated by the stochastic nature of the domain—both *supply* (number of impressions or CTs) and *demand* (future bids) are uncertain—which suggests the need for online allocation.

To address these issues, we introduced the idea of an *optimize-and-dispatch* architecture [6] where an optimizer is run only every so often and it parameterizes a dispatcher that operates in real time. The optimizer can be run at fixed intervals, or based on any other trigger conditions, such as supply or demand significantly deviating from their projections. The framework can, in principle, handle any forms of expressive preferences as inputs, and we discuss several forms of expressiveness that are important in ad auctions, but which prior ad auction mechanisms inherently cannot support.

We recently implemented these ideas [4]. We model the problem as a Markov decision process (MDP), whose solution is approximated in several ways. First we perform optimization only periodically. Following the general optimize-and-dispatch framework, our optimization generates an on-line *dispatch policy* that assigns ad channels to advertisers in real-time. Our dispatch policies use the fractional assignment of (dynamically defined) channels to specific contracts. To approximate the optimization itself, we consider two approaches. The first is deterministic optimization using expectations of all random variables and exploiting our combinatorial optimization technology for winner determination in expressive markets [7]. We propose a second, sample-based approach derived from van Hentenryck and Bent's [5] online model for stochastic optimization—but with novel adaptations to a continuous decision space. This approach leverages the deterministic winner determination technology, applying it to multiple possible future scenarios in order to form a

[1]Here we use the word "channel" totally differently than in the "channel-based" mechanisms discussed earlier in this abstract. Here, each "channel" is a subset of supply such that no bid distinguishes between different forms of supply within the channel.

dispatch policy. In both cases, periodic reoptimization is used to overcome the approximate nature of the methods. Our experiments demonstrate the benefits of expressive bidding for ad auctions over various per-item strategies, and the value of our stochastic optimization techniques.

References

[1] Chris Anderson. *The Long Tail: Why the Future of Business Is Selling Less of More.* Hyperion, July 2006.

[2] Michael Benisch, Norman Sadeh, and Tuomas Sandholm. The cost of inexpressiveness in advertisement auctions. In *Proceedings of the Fourth Workshop on Ad Auctions*, 2008.

[3] Michael Benisch, Norman Sadeh, and Tuomas Sandholm. A theory of expressiveness in mechanisms. In *Proceedings of National Conference on Artificial Intelligence (AAAI)*, 2008.

[4] Craig Boutilier, David Parkes, Tuomas Sandholm, and William Walsh. Expressive banner ad auctions and model-based online optimization for clearing. In *Proceedings of National Conference on Artificial Intelligence (AAAI)*, 2008.

[5] Pascal Van Hentenryck and Russell Bent. *Online Stochastic Combinatorial Optimization.* MIT Press, 2006.

[6] David Parkes and Tuomas Sandholm. Optimize-and-dispatch architecture for expressive ad auctions. In *First Workshop on Sponsored Search Auctions, at the ACM Conference on Electronic Commerce*, Vancouver, BC, Canada, June 2005.

[7] Tuomas Sandholm. Expressive commerce and its application to sourcing: How we conducted $35 billion of generalized combinatorial auctions. *AI Magazine*, 28(3):45–58, 2007.

[8] Tuomas Sandholm, David Levine, Michael Concordia, Paul Martyn, Rick Hughes, Jim Jacobs, and Dennis Begg. Changing the game in strategic sourcing at Procter & Gamble: Expressive competition enabled by optimization. *Interfaces*, 36(1):55–68, 2006.

[9] Tuomas Sandholm and Subhash Suri. Side constraints and non-price attributes in markets. *Games and Economic Behavior*, 55:321–330, 2006. Extended version in IJCAI-2001 Workshop on Distributed Constraint Reasoning.

Tuomas Sandholm
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213, USA
Email: `sandholm@cs.cmu.edu`

Ranking, Trust, and Recommendation Systems: An Axiomatic Approach

Moshe Tennenholtz

In the classical theory of social choice, a theory developed by game-theorists and theoretical economists, we consider a set of agents (voters) and a set of alternatives. Each agent ranks the alternatives, and the major aim is to find a good way to aggregate the individual preferences into a social preference. The major tool offered in this theory is the axiomatic approach: study properties (termed axioms) that characterize particular aggregation rules, and analyze whether particular desired properties can be simultaneously satisfied. In a ranking system [1] the set of voters and the set of alternatives coincide, e.g. they are both the pages in the web; in this case the links among pages are interpreted as votes: pages that page p links to are preferable by page p to pages it does not link to; the problem of preference aggregation becomes the problem of page ranking. Trust systems are personalized ranking systems [3] where the ranking is done for (and from the perspective of) each individual agent. Here the idea is to see how to rank agents from the perspective of a particular agent/user, based on the trust network generated by the votes. In a trust-based recommendation system the agents also express opinions about external topics, and a user who has not expressed an opinion should be recommended one based on the opinions of others and the trust network [6]. Hence, we get a sequence of very interesting settings, extending upon classical social choice, where the axiomatic approach can be used.

On the practical side, ranking, reputation, recommendation, and trust systems have become essential ingredients of web-based multi-agent systems (e.g. [9, 13, 7, 14, 8]). These systems aggregate agents' reviews of products and services, and of each other, into valuable information. Notable commercial examples include Amazon and E-Bay's recommendation and reputation systems (e.g. [12]), Google's page ranking system [11], and the Epinions web of trust/reputation system (e.g. [10]). Our work shows that an extremely powerful way for the study and design of such systems is the axiomatic approach, extending upon the classical theory of social choice. In this talk we discuss some representative results of our work [14, 1, 2, 4, 5, 3, 6].

References

[1] A. Altman and M. Tennenholtz. On the axiomatic foundations of ranking systems. In *Proc. 19th International Joint Conference on Artificial Intelligence*, pages 917–922, 2005.

[2] A. Altman and M. Tennenholtz. Ranking systems: the pagerank axioms. In *EC '05: Proceedings of the 6th ACM conference on Electronic commerce*, pages 1–8, New York, NY, USA, 2005. ACM Press.

[3] A. Altman and M. Tennenholtz. An axiomatic approach to personalized ranking systems. In *Proc. 20th International Joint Conference on Artificial Intelligence*, 2007.

[4] A. Altman and M. Tennenholtz. Quantifying incentive compatibility of ranking systems. In *Proc. of AAAI-06*, 2006.

[5] A. Altman and M. Tennenholtz. Incentive compatible ranking systems. In *Proceedings of the 2007 International Conference on Autonomous Agents and Multiagent Systems (AAMAS-07)*, 2007.

[6] R. Andersen, C. Borgs, J. Chayes, U. Feige, A. Flaxman, A. Kalai, V. Mirrokni, and M. Tennenholtz. Trust-Based Recommendation Systems: an Axiomatic Approach. In *Proceedings of WWW-08*, 2008.

[7] Y. Bakos and C. N. Dellarocas. Cooperation without enforcement? a comparative analysis of litigation and online reputation as quality assurance mechanisms. MIT Sloan School of Management Working Paper No. 4295-03, 2003.

[8] R. Dash, S. Ramchurn, and N. Jennings. Trust-based mechanism design. In *Proceedings of the Third International Joint Conference on Autonomous Agents and MultiAgent Systems*, pages 748–755, 2004.

[9] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[10] P. Massa and P. Avesani. Controversial users demand local trust metrics: An experimental study on epinions.com community. In *Proc. of AAAI-05*, pages 121–126, 2005.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report, Stanford University, 1998.

[12] P. Resnick and R. Zeckhauser. Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. Working Paper for the NBER workshop on empirical studies of electronic commerce, 2001.

[13] P. Resnick, R. Zeckhauser, R. Friedman, and E. Kuwabara. Reputation systems. *Communications of the ACM*, 43(12):45–48, 2000.

[14] M. Tennenholtz. Reputation systems: An axiomatic approach. In *Proceedings of the 20th conference on uncertainity in Artificial Intelligence (UAI-04)*, 2004.

Moshe Tennenholtz
Technion – Israel Institute of Technology
Haifa 32000, Israel
Email: `moshe@tie.technion.ac.il`

Lorenz Rankings of Rules for the Adjudication of Conflicting Claims

William Thomson

For the problem of adjudicating conflicting claims (O'Neill, 1982; for a survey, see Thomson, 2003), we offer simple criteria to compare rules on the basis of the Lorenz order. They exploit several recently developed techniques to structure the space of rules.

The first results concern the family of "ICI rules" (Thomson, 2008, forthcoming). This family contains the constrained equal awards (Maimonides, 12th Century), constrained equal losses (Maimonides, 12th Century), Talmud (Aumann and Maschler, 1985), and minimal overlap (Ibn Ezra, 12th Centrury; O'Neill, 1982) rules. We obtain a condition relating the parameters associated with two rules in the family guaranteeing that one Lorenz dominates the other. We prove parallel results for a second family (CIC family, Thomson, 2008, forthcoming), which is obtained from the first one by exchanging, for each problem, how well agents with relatively larger claims are treated as compared to agents with relatively smaller claims. This second family also contains the constrained equal awards and constrained equal losses rules.

The next results concern the family of "consistent" rules (Young, 1987). A rule is consistent if the recommendation it makes for each problem is never contradicted by the recommendation it makes for each reduced problem obtained by imagining some claimants leaving the scene with their awards and reassessing the situation at that point. The main result here is the identification of circumstances under which the Lorenz order is "lifted by consistency" from the two-claimant case to arbitrarily many claimants. (The concept of lifting is adapted from one developed for properties of rules by Hokari and Thomson, 2008, forthcoming.) This means that if two rules are consistent and one of them Lorenz dominates the other in the two-claimant case, this domination extends to arbitrarily many claimants.

Finally, we exploit the notion of an operator on the space of rules (Thomson and Yeh, 2008, forthcoming). An operator is a mapping that associates with each rule another one. The operators we consider are the duality operator, the claims truncation operator, and the attribution of minimal right operator. We also consider the operator that associates with each rule and each list of non-negative weights for them adding up to one, their weighted average. An operator "preserves an order" if whenever a rule Lorenz dominates another one, this domination extends to the two rules obtained by subjecting them to the operator. We identify circumstances under which certain operators preserve the Lorenz order (or reverse it), and circumstances under which a rule can be Lorenz compared to the rule obtained by subjecting it to the operator.

William Thomson
Department of Economics
University of Rochester
Rochester, NY 14627, USA
Email: wth2@troi.cc.rochester.edu

A fair payoff distribution for myopic rational agents

Stéphane Airiau and Sandip Sen

Abstract

We consider the case of self-interested agents that are willing to form coalitions for increasing their individual rewards. We assume that each agent gets an individual payoff which depends on the coalition structure (CS) formed. We consider a CS to be stable if no individual agent has an incentive to change coalition from this CS. Stability is a desirable property of a CS: if agents form a stable CS, they do not spend further time and effort in selecting or changing CSs. When no stable CSs exist, rational agents will be changing coalitions forever unless some agents accept suboptimal results. When stable CSs exist, they may not be unique, and choosing one over the other will give an unfair advantage to some agents. In addition, it may not be possible to reach a stable CS from any CS using a sequence of myopic rational actions. We provide a payoff distribution scheme that is based on the expected utility of a rational myopic agent (an agent that changes coalitions to maximize immediate reward) given a probability distribution over the initial CS. To compute this expected utility, we model the coalition formation problem with a Markov chain. Agents share the utility from a social welfare maximizing CS proportionally to the expected utility of the agents, which guarantees that agents receive at least as much as their expected utility from myopic behavior. This ensures sufficient incentives for the agents to use our protocol.

1 Introduction

In the literature on coalition formation, valuation functions are typically defined only over a coalition, and the agents need to decide or negotiate a payoff distribution. We are interested in cases where the payoff distribution is defined for each partition of the agents into coalition structures (CS): each agent knows its payoff for any CS. This model corresponds to the hedonic aspect of coalition formation [1, 2, 4, 7] where the payoff of an agent, not the value of a coalition, depends only on the members of its coalition. We can view this assumption from two perspectives. The first perspective is that the environment provides a payoff to each agent. This can happen when the agents' individual goals are different but correlated and the CSs have different effects on different agents' performance. This formulation can model a community of agents that help each other improve their respective private utilities: each agent obtains a private utility which can be boosted with the help of other agents in the community. Another example is that of firms forming coalitions in a supply chain domain: each firm in a coalition provides preferred rates or discounts for its services to other members of its coalition. The benefit of each member of the coalition depends on the behaviors of other firms. Each firm in the coalition is autonomous: each sells and buys goods, and makes its own profit or loss. Note that firms still benefit from being in a coalition but the benefit varies from firm to firm in any given CS. The second perspective is to consider that the payoff distribution has already been computed using a stability criteria, e.g., the Kernel [5]. Given a CS and a valuation function, it is always possible to compute a Kernel-stable payoff distribution. Let us consider two different CSs with associated Kernel-stable payoff distributions. In both cases, the payoff distribution is stable, but an agent may prefer to form the first CS when another agent would prefer the second: even if agents are using a stable payoff distribution, agents may still have incentive to change CS.

In the papers related to hedonic coalition formation [1, 2, 4, 7], one assumption is that there is no transfer of utility. Under these assumptions it is known that the core or the set of Nash-stable equilibria may be empty. In particular, the personal goals of the agents may be conflicting, there may not be any CS that satisfies all the agents at the same time: for each CS, at least one agent may have an incentive to change coalitions. Some research deals with the search of conditions for the existence of stable coalition structures [1, 2, 4], but we want to provide a solution even when no stable CSs exist. The compromise we propose is based on allowing transfer of utility to make a CS stable.

From a societal point of view, we also want the society to perform well as a whole, hence our mechanism selects a social welfare maximizing CS. The computation of the side-payments to stabilize the CS is based on the expected utility of myopic and rational agents (i.e., agents that change coalition to maximize their immediate payoff). The computation of the expected utility uses the analysis of a Markov chain where a state of the chain corresponds to a CS and a transition corresponds to the will of an agent to change coalition. The analysis of the chain differentiate the transient states from the ergodic states[1], the latest corresponds to the Sink equilibria in [9] for games in normal forms: myopic rational agents are bound to be trapped in a set of ergodic states. The expected utility of the agent is a weighted average of the utility over the ergodic states. We view the expected utility as a means to weight the importance of an agent in the coalition formation process. We share the utility of a social welfare maximizing CS proportionally to this expected utility. Under the assumption that the initial CS is chosen at random, we show that each agent is better off following our protocol.

Most current studies on coalition formation in the multiagent community assume known valuation functions that estimate the worth of a coalition and where the valuation of any coalition is independent of the other coalitions present in the population [10, 14, 15]. However, this may not always be appropriate. In situations when the population of agents is competing for a resource or a niche, e.g., in electronic supply chains, the valuation of a coalition depends on the organization of the other agents. More generally, the presence of shared resources (if a coalition uses some resource, they will not be available to other coalitions) or conflicting goals (non-members can move the world farther from a coalition's goal state) [13] makes a valuation function depend on CS. We are especially interested in studying those situations where the worth of a coalition depends on the other coalitions that are present in the population [6, 12]. Our approach can also be used in that context as shown by our empirical example.

The paper is organized as follows. In Section 2, we present the coalition framework and the existing stability concepts for coalition formation when in the non transferable utility case. In Section 3, we show how to build a Markov chain that models the coalition formation process, how to use it to compute the expected utility. Finally in Section 4, we present and discuss our proposed solution. We conclude and discuss future work in Section 5.

2 Coalition Framework

2.1 Problem Description

We consider a set N of n agents; N is also known as the *g*rand coalition. A coalition structure (CS) $s = \{\mathcal{S}_1, \cdots, \mathcal{S}_m\}$ is a partition of N, where $\mathcal{S}_i$ is the i^{th} coalition of agents with $\cup_{i\in[1..m]}\mathcal{S}_i = N$ and $i \neq j \Rightarrow \mathcal{S}_i \cap \mathcal{S}_j = \emptyset$. $\mathscr{S}$ is the set of all CSs. The coalition of agent i in s is noted as $s(i)$. We consider that an agent i has a preference order $\succsim_i$ over $\mathscr{S}$

[1]Ergodic states are states that the chain will keep coming back to, whereas transient states are states that the chain will eventually leave to never visit again.

and for a CS s, an agent i has a valuation $v_i(s)$. These assumptions have two consequences:

- Each agent has a private utility which depends on the other agents present in the coalition, as for hedonic coalition formation [1, 4, 2, 7]. Coalitions do not always receive a reward as a whole: each agent has a private cost and benefit which depends on the organization of the agents. Members of a coalition help each other, which can globally reduce the cost or increase the private benefit of each member. For example, soccer players have a private utility, or satisfaction, which depends on the other members of the team.

- Unlike in the hedonic coalition formation case, we are working in the more general case where the valuation of a coalition depends on the other coalitions present in the population. For an agent i such that $i \in \mathcal{C}$ and two CSs s_1 and s_2 such that $\mathcal{C} \in s_1$ and $\mathcal{C} \in s_2$, it is possible that $u_i(s_1) \neq u_i(s_2)$. In our soccer example, the satisfaction of a player in a team playing a league may also depend on how the remaining players are dispatched in the other team, for example, he may prefer that the best players are put in different teams than put altogether in a "dream team". A more generic example involves agents competing for an environmental niche. The payoff of a coalition may be higher when the competitors work alone than when the competitors also decide to team together to form a more competitive group. Ray and Vohra [12] consider this problem and propose a protocol where agents propose a coalition and a distribution of the coalitions' worth. Other agents can accept or reject the proposition. One issue is that, when proposing a coalition, an agent does not know which CS will ultimately form. Hence, the payoff distribution proposed by an agent is conditioned on the CS that is finally selected. Ray and Vohra consider that the agents' offer contains a payoff distribution for each possible CS, which is not realistic for large populations. But such elaborate offers allow them to show the existence of an equilibrium.

We further assume that there is no coordinated change of coalitions; one agent at a time can change coalition. This assumption prevents uncertainties about the state of the CS. For example, let agents i, j and k form singleton coalitions. At this point, agent i would like to join agent k, and agent j would like to join agent i, but neither i or j would like to form the grand coalition. If we allowed simultaneous moves, the resulting state would be unclear. The grand coalition may be formed though it was not the intent of agent i or j. The resultant CS could also be $\{\{i, k\}, \{j\}\}$ where agent i joined agent k, and agent j tried to join the coalition of agent i, but ended up joining an empty coalition. It could also be $\{\{i, j\}, \{k\}\}$ where agent j joined agent i, and agent k refused that both agent i and j joined it at the same time. We avoid such ambiguities with this assumption.

Finally, we assume that agents are myopic and rational, and members of a coalition accepts a new member only when all members agree. After a change of CS, it is possible, if not likely, that another agent changes its coalition, leading to a different CS. As it is computationally expensive to perform multi-steps look ahead because of the large state space, we consider myopic agents that change coalition to maximize their immediate reward. We believe it is reasonable to assume that current members can control when other agents can join a coalition. Moreover, it would not be myopic rational for a member i to accept a new agent if this meant a payoff loss for i. Hence, we also assume that all members of a coalition must agree to accept a new member and, if some member i refuses, we will say that agent i vetoes the transition. We also make the implicit assumption that members of a coalition cannot prevent a member to leave, even if some of the remaining members lose utility.

2.2 Stability Concepts

We first start by giving the definition of stability concepts in the non-transferable utility case when the value function depends only on the members of the coalition [4]. In the following, $\succsim_i$ denotes a preference order over coalitions.

Definition 2.1. *A coalition structure s is* ***core stable*** *iff* $\nexists C \subset N \mid \forall i \in C, C \succ_i s(i)$.

Definition 2.2. *A coalition structure s is* ***Nash stable*** $(\forall i \in N)\ (\forall C \in s \cup \{\emptyset\})\ s(i) \succsim_i C \cup \{i\}$

Definition 2.3. *A coalition structure s is* ***individually stable*** *iff* $(\nexists i \in N)\ (\nexists C \in s \cup \{\emptyset\}) \mid (C \cup \{i\} \succ_i s(i))$ *and* $(\forall j \in C,\ C \cup \{i\} \succsim_j C)$

Definition 2.4. *A coalition structure s is* ***contractually individually stable*** *iff* $(\nexists i \in N)\ (\nexists C \in s \cup \{\emptyset\}) \mid (C \cup \{i\} \succ_i s(i))$ *and* $(\forall j \in C,\ C \cup \{i\} \succsim_j C)$ *and* $(\forall j \in s(i) \setminus \{i\},\ s(i) \setminus \{i\} \succsim_j s(i))$

If a CS is core stable, no subset of agents has incentive to leave their respective coalition to form a new one. In a Nash stable CS s, no single agent i has an incentive to leave its coalition $s(i)$ to join an existing coalition in s or create the singleton coalition $\{i\}$. The two other criteria add a constraint on the members of the coalition joined or left by the agent. For an individually stable CS, there is no agent that can change coalition from s(i) to $C \in (s \setminus s(i)) \cup \{\emptyset\}$ yielding better payoff for itself, and the members of C should not lose utility. The contractually individual stability in addition requires that the members of s(i), the coalition left by i, should not lose utility.

The definition of Nash, individually and contractually individually stability can easily be extended to the case where the value of a coalition depends on the CS. Another criterion for a rational agent to be a member of a coalition is individual rationality [6]: an agent i would consider joining a coalition only when it is beneficial for itself. The agent compares the situation when it is on its own and when it is a member of a coalition. However, the payoff the agent gets when it is by itself depends on the CS. The minimum payoff that agent i can guarantee on its own is $r_i = \min_{s \in \mathscr{S}, \{i\} \in s} v_i(s)$ [6] (the minimum is over all the CSs where agent i forms a coalition on its own). An agent is individual rational when its payoff in a coalition with other agents is greater than the minimum payoff it can get on its own.

For some coalition formation problem, it is possible that no CS satisfies any of these stability criteria. Satisfying the individually or contractually individually stability criteria may depend on the protocol used by the agents to form coalition. For example, an academic can freely leaves its department to join a new one, provided that no member of the new department will suffer from its presence. In some cases, the coalition left is allowed to demand compensation. For example, as pointed out in [7], a player of a soccer team can join another club, but its former club can receive a compensation for the transfer. In the following, we will only assume that members of a coalition can veto the entrance of new agent in their coalition. Hence, we consider as our main stability criterion the individually stability.

2.3 Graphical representation

We can represent the relation $\succsim$ by a **preference graph of the coalition formation process**: each node is a CS, and there exists an edge from node S to node T when $\exists i \in N \mid T \succ_i S$. The **transition graph of the coalition formation process** is a directed graph where the nodes represent the CSs, and edges are valid transitions between two CSs. A transition from node s to node t is valid when

- $\exists i \in N$, $\exists C \in (t \setminus s(i)) \cup \{\emptyset\} \mid t = (s \setminus s(i) \setminus C) \cup (C \cup \{i\})$ and $t \succ s$. In other words, there is an agent i that is better off leaving its coalition $s(i)$ to either join an existing coalition in s or to form a singleton coalition.
- and $\forall j \in C$, $t \succsim_j s$, i.e., this transition is not vetoed by the members of the coalition C joined by i (of course, i is always allowed to form a singleton coalition).

Incidentally, another agent j may also prefer t over s (for example, when i moves to an existing coalition C, all agents in C may benefit). Hence, a transition may be beneficial for more than one agent. However, only the agent that changes the coalition can induce the transition. Even if it is beneficial for members of C, C's members cannot force i to leave its current coalition to join them (this action would be considered to be a group action whereas in our model, we consider only individual actions). In the case where two agents i and j that were previously forming singleton coalitions now form a coalition of two agents in the new CS, it may be difficult to interpret which agent induced the transition: as it is beneficial for both agents, an interpretation of the transition can be that agent i joins the coalition $\{j\}$ or vice versa. Our interpretation is that both agents are responsible for this transition. Hence we make an exception for this case.

Since we assume that agents are myopically rational, for a given CS, each agent will only choose the transition that yields the maximum immediate payoff gain over all its possible legal moves. For each state, there can then be at most n outgoing edges, one for each of the n agents (this happens when every agent prefer another CS over the current one). This prunes the number of transitions from the preference graph to the transition graph.

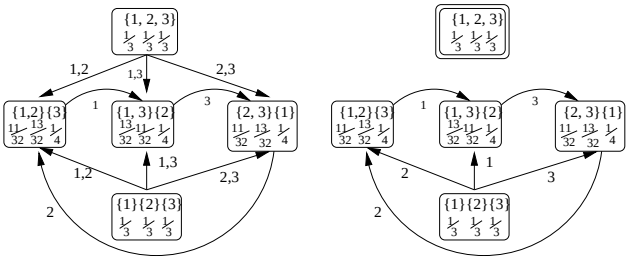
Property 1. *A CS s is individually stable iff there is no outgoing edge from state s in the transition graph of the coalition formation.*

The proof is obvious given the definition of the transition graph. In Figure 1(a), we present an example with three agents where the payoff of an agent is shown below its label in a coalition. In this example, no CS is core or Nash stable. However, $\{\{1, 2, 3\}\}$ is individually stable. However, if the agents start from the bottom of the lattice (where each agent forms a singleton coalition) or any other CS in the mid level, the agents will be trapped in a cycle: for each CS in the mid-level, one agent benefits from leaving its coalition in that CS to join the singleton agent. We present a different scenario in Figure 1(b): the CS $\{\{0\}\{1, 2\}\}$ is Nash stable, core stable (and hence individually stable), and the grand coalition is individually stable. From any CS, it is possible to reach an individually stable CS.

3 A Markov Chain model

A myopic rational agent will change coalitions if it can immediately gain utility by doing so. In this paper, we assume that the valuation is common knowledge. It is therefore possible to build and analyze the transition graph. Given the assumption that only one agent at a time can change coalition, we are now in position to estimate the probability of transition between any two CSs. For each outgoing edge e from CS s, the probability of making this transition is either

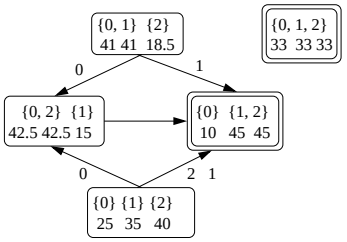
- $\frac{1}{o(s)}$, where $o(s)$ is the out degree of a node, i.e., the number of agents that want to change from s.
- $\frac{2}{o(s)}$ when two agents i and j that are each forming a singleton coalition merge to form the two-agent coalition $\{i, j\}$ and it is the best choice for both i and j.



Preference graph Transition model

The valuation function has one individually stable CS and contains a cycle.

(a) Coalition formation with a cycle



(b) Transition model with two individually stable CSs

Figure 1: Example of Coalition Formation problem (double boxed CS are individually stable)

$$
\begin{array}{c}
\\
\{1,2,3\} \\
\{1,2\}\{3\} \\
\{1,3\}\{2\} \\
\{2,3\}\{1\} \\
\{1\}\{2\}\{3\}
\end{array}
\begin{array}{c}
\begin{array}{ccccc}
\{1,2,3\} & \{1,2\}\{3\} & \{1,3\}\{2\} & \{2,3\}\{1\} & \{1\}\{2\}\{3\}
\end{array} \\
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0
\end{pmatrix}
\end{array}
$$

Table 1: Transition Matrices for Figure 1(a)

As the probability of a transition does not depend upon the prior states of the population, the Markov assumption is verified. We have now completely defined a Markov chain. From the above specified transition model, we can construct the transition matrix P of the Markov chain. The size of the matrix is $\mathscr{B}(n) \times \mathscr{B}(n)$, where n is the number of agents and $\mathscr{B}(n)$ is the Bell function. The dimension of the matrix can be quite large, however, the matrix is sparse: for each row of the matrix, there can be only up to n positive entries[2]. In Table 1, we present the transition matrix for the example of Figure 1(a).

As agents change from one CS to another, the chain moves from one state to another. A state of a Markov chain is either transient or ergodic: ergodic states are states that the chain will keep coming back to, whereas transient states are states that the chain will eventually leave to never visit again. In the long term, the chain will be in one of the ergodic states. The ergodic states form multiple strongly connected components. If the size of such a strongly connected component is one, it means that the corresponding CS is individually stable (it may also be core or Nash stable, but not necessarily). The study of the Markov chain will tell us, given a probability distribution over the initial state, the probability to reach each strongly connected component, and, once reached, what is the proportion of time spent in each ergodic states. Hence, the value of the expected utility is an average over the

[2] $\mathscr{S}$ can be represented by a lattice where each CS at a given level of the lattice contains the same number of coalitions. For each level i in the lattice, an agent has at most i actions: joining one of the existing $i-1$ coalitions and forming a singleton coalition if it is not already forming one. As there are n levels, the maximum number of transitions from a CS is n.

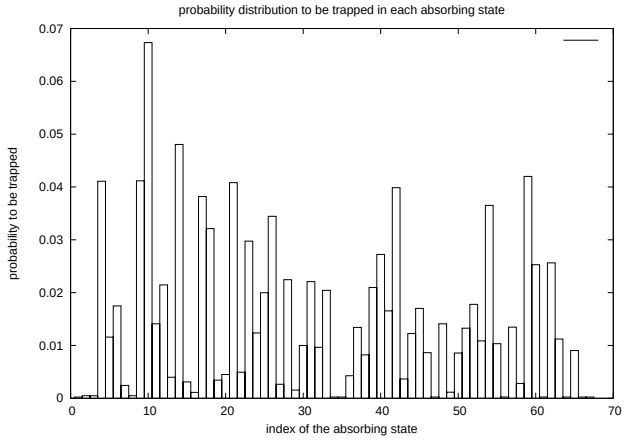


Figure 2: Example of the ART domain: probability to be in an ergodic state

possible stable CSs, and the CSs that are parts of some cycle. More formally, let $\mathscr{E}$ be the set of ergodic states of the Markov chain. For each strongly connected component $X \subset \mathscr{E}$, we compute the probability $p_X$to reach X, and then for each state s of X, we compute the fraction of time p_s spent in s in the limit (the chain may visit a state in X more often then another). If a CS s is at least individually stable, then the size of the corresponding component is one, and $p_s = 1$. For each ergodic state $s \in \mathscr{E}$, let $X(s)$ be the strongly connected component of s. The expected utility $E(v_i)$ is then $E(v_i) = \sum_{s \in \mathscr{E}} p_{X(s)} \cdot p_s \cdot v_i(s)$.

In Figure 2, we present an example issue from the Agent Reputation and Trust testbed [8]. In the testbed, agents provide appraisals about artifacts and compete for a pool of clients. To improve their appraisals, agents can ask other agents for appraisals for artifacts and reputation of other agents. We consider collusion of agents: agents can form a coalition where members provide their truthful appraisals, which benefits all members. In a domain with 8 agents, we computed the valuation function and the associated Markov chain for a particular instance, and the outcome is presented in Figure 2. In that instance, the Markov chain contains 4,140 CSs, 26,641 transitions, 62 stable CSs and 5 additional ergodic states which correspond to some strongly connected components.

4 A Fair Payoff Distribution for Myopic Rational Agent

It is possible that some coalition formation problem do not have any stable CS. To operate efficiently, we require that the agents remain in a CS. We propose that the agent forms a CS $s^\star$ that maximizes social welfare. However, $s^\star$ may not be stable, hence we want to share the utility $u^\star$ of $s^\star$ that provides the agent an incentive to stay in that CS.

The utility function, as a whole, tells how good the agent is. A first candidate is to share $u^\star$ proportionally to the average utility over all the CSs. This assumes that each CS is equally important and we believe it is not so. Another candidate is to consider an average over the stable CSs. However, such stable CSs may not always exist, and even if they do, there may not be a path allowing to reach a stable CS (as in the example of Figure 1(a)). If we assume any CS is likely to be the initial CS, we can compute an expected utility when the agents are myopic, rational, and when members of a coalition can veto the entrance of new members. The expected utility is a great metric to determine and compare the strength of each agent in the coalition formation process. We will show that the payoff obtained is

at least its expected utility, which is a sufficient incentive for using our proposed payoff distribution.

4.1 Choice of Final Payoff Distribution and Corresponding CS

The expected value $E(v_i)$we computed using the Markov chain assumes that the initial CS is chosen uniformly over $\mathscr{S}$, in other word, it is no biased by the initial CS. $E(v_i)$ reflects the utility that agent i receives on average when all agents are myopically rational. We consider that this value represents the strength of an agent given the valuation function. Agents with high $E(v_i)$ should obtain a larger payoff than agents with lower $E(v_i)$.

To be used in a real world application, it is not desirable to have agents continuously change coalitions: agents should form a stable CS and have no incentive to further change coalition. To maximize the agents' payoff, we choose as the final CS $s^\star$ one of the CSs that maximizes social welfare. This CS may not be a stable, but it guarantees maximal total payoff to the agents. As we view the expected utility value as a measure of the strength of each agent, we propose a distribution of $v(s^\star)$ to all agents proportional to the expected payoff of the agents, i.e., we prescribe the payoff to agent i to be

$$u_i = \frac{E(v_i)}{\sum_{j\in N} E(v_j)} v(s^\star).$$

Note that this value is guaranteed to be at least as good as $E(v_i)$, as shown by Property 2. So, when agents share the payoff we propose, they are guaranteed to have at least the expected value when they were changing coalitions to maximize their immediate reward, and in general, they may get more. In addition, the payoff distribution is Pareto Optimal as we share the value of a social welfare maximizing CS (if an agent gets more utility, at least another agent must lose some). We believe that these incentives are sufficient for the agents to accept our proposed value. Not only is the payoff distribution fair, as the share of utility the agents receive is proportional to their expected utility over the chain, but the outcome is also efficient as it maximizes social welfare.

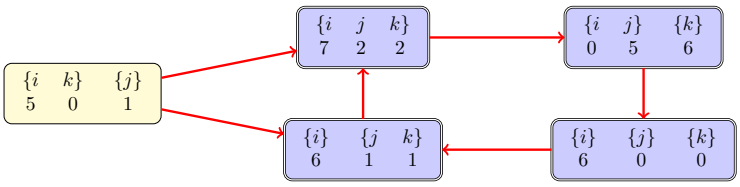
Property 2. *$u_i = \frac{E(v_i)}{\sum_{j\in N} E(v_j)} v(s^\star) \geq E(v_i)$, i.e., the payoff of an agent is at least as good as the expected utility that an agent would get on average if the agents are myopically rational.*

Proof. Let $\mathscr{E}$ denote the set of the ergodic states of a Markov chain. For player i, the expected payoff is a weighted average over the ergodic states: $E(v_i) = \sum_{s\in\mathscr{E}} \alpha_s v_i(s)$, where α_s is the weight of the ergodic state s and we have $\sum_{s\in\mathscr{E}} \alpha_s = 1$. The transient states are only used to determine the probability of leading to one of the ergodic sets: the α_s's are determined by the transient and the ergodic states (when there is a cycle or a regular sub-chain).

$$\begin{aligned}
\forall s,\ v(s) &\leq v(s^\star) \\
\forall s,\ \alpha_s \cdot v(s) &\leq \alpha_s \cdot v(s^\star) \text{ as } \alpha_s \geq 0 \\
\textstyle\sum_{s\in\mathscr{E}} \alpha_s \cdot v(s) &\leq \textstyle\sum_{s\in\mathscr{E}} \alpha_s \cdot v(s^\star) \\
\textstyle\sum_{s\in\mathscr{E}} \alpha_s \cdot v(s) &\leq v(s^\star) \cdot \textstyle\sum_{s\in\mathscr{E}} \alpha_s \\
\textstyle\sum_{s\in\mathscr{E}} \alpha_s \cdot v(s) &\leq v(s^\star), \text{ as } \textstyle\sum_{s\in\mathscr{E}} \alpha_s = 1 \\
\textstyle\sum_{s\in\mathscr{E}} \alpha_s \sum_{j\in N} v_j(s) &\leq v(s^\star) \\
\textstyle\sum_{j\in N} \sum_{s\in\mathscr{E}} \alpha_s v_j(s) &\leq v(s^\star) \\
\textstyle\sum_{j\in N} E(v_j) &\leq v(s^\star) \\
E(v_i) &\leq \tfrac{E(v_i)}{\sum_{j\in N} E(v_j)} v(s^\star) \text{ as } E(v_i) \geq 0.
\end{aligned}$$

□

Another important question is to determine whether the payoff distribution v_i is individually rational: is an agent guaranteed to get as much as when an agent is forming a singleton coalition? The minimum payoff an agent can guarantee for itself is $r_i = \min_{s \in \mathscr{S},\, \{i\} \in s} v_i(s)$. For example, consider the three-agent example in Figure 3. The value obtained by i is $\frac{209}{36} = 5.806$, which is lower than 6, the minimum payoff that agent i receives when it forms a singleton coalition. This means that the payoff obtained by an agent in a coalition from our protocol is less than the worst payoff obtained by the agent when it forms a singleton coalition. Although possible in the general case, this may not be likely in practice: the worst case scenario for an agent should be when it forms a singleton coalition and when all other agents in the population try to minimize its payoff. As shown by Property 3, if the worst payoff for an agent occurs when it is forming a singleton coalition, our protocol is individual rational.



There is a cycle with 4 states, hence, the proportion spent in each state is $\frac{1}{4}$. The value of the optimal CS is 11. The minimum value of agent i when it is in a singleton coalition is 6.

$E(v_i) = \frac{1}{4}(7+0+6+6) = 4.75$ $\quad v_i = \frac{4.75}{4.75+2+2.25} \cdot 11 = 5.8056 < 6 = \frac{216}{36}$

$E(v_j) = \frac{1}{4}(2+5+0+1) = 2$ $\quad v_j = \frac{2}{4.75+2+2.25} \cdot 11 = 2.4444 > 0$

$E(v_k) = \frac{1}{4}(2+6+0+1) = 2.25$ $\quad v_k = \frac{2.25}{4.75+2+2.25} \cdot 11 = 2.75 > 0$

Figure 3: Case where the protocol is not individual rational: i's payoff is lower than r_i, i's minimum payoff when it forms a singleton.

Property 3. *If* $(\forall s \in \mathscr{S})\ v_i(s) \geq r_i = \min_{s \in \mathscr{S} | \{i\} \in s}$, *then* $u_i \geq r_i$, *i.e., the payoff distribution* u_i *is individually rational.*

Proof. The hypothesis $\forall s \in \mathscr{S}$, $v_i(s) \geq r_i$ means that for any CS, the valuation of agent i is at least equal to i's minimum valuation when it forms a singleton coalition, i.e., the payoff of an agent in a coalition with at least another agent should be at least the minimum payoff the agent receives when it is on its own in a singleton coalition. Hence, we have $\sum_{s \in \mathscr{E}} \alpha_s v_i(s) \geq \sum_{s \in \mathscr{E}} \alpha_s r_i$, and then $E(v_i) \geq r_i$ as $\sum_{s \in \mathscr{E}} \alpha_s = 1$. From Proposition 2, we have $u_i \geq E(v_i) \geq r_i$. □

4.2 Computational Complexity of the centralized algorithm

We now consider the complexity of computing the payoff distribution if a centralized entity was used. To compute the canonical form of a stochastic matrix, we first need to compute the communication classes of the matrix and this operation is polynomial in the size of the matrix ($O(\mathscr{B}(n)^2)$). Then, to determine the canonical form of the matrix, we need to find the permutation matrix, which can also be done in quadratic time, hence in $O(\mathscr{B}(n)^2)$. To compute the limit behavior of the Markov chain, either a matrix has to be inverted (which can be done in $O(\mathscr{B}(n)^3)$, or a linear system needs to be solved (iterative methods can also be used here). The complexity is then $O(\mathscr{B}(n)^3)$. The fact that the matrix is sparse should allow for faster computation. The search of the optimal CS is $O(\mathscr{B}(n))$ if the brute force method is applied. As we consider valuation function that depends on CS, we cannot use the faster algorithm in [11]. The computation of the side-payments and the execution of the payments has linear complexity. Hence, the complexity of the protocol is $O(\mathscr{B}(n)^3)$.

4.3 Experiments with Random Valuation Function

We now experiment with random valuation functions. The valuation of a coalition $\mathcal{C}$ for a particular CS is drawn from a uniform distribution in $[0, \mathcal{C}]$. Using this distribution, it is on average better to have coalitions containing many agents, but the valuation function is not superadditive. The valuation of each member of $\mathcal{C}$ is distributed randomly: each member $i \in \mathcal{C}$ receives $w_i \cdot v(\mathcal{C})$ with z_i drawn from a uniform distribution in $[0, 1]$ and $w_i = \frac{z_i}{\sum_{j \in \mathcal{C}} z_i}$. We now present the result of a particular valuation function with 6 agents where the number of CSs is 203. The associated Markov chain has 54 transient states and 149 ergodic states. The associated transition matrix of size $203 \times 203 = 41209$ has only 735 positive entries, the matrix is quite sparse. The CS with maximal social welfare is not individually stable. The value of the agents are shown in Table 4.3: the second column (avg) represents the average payoff of an agent over all CSs, the third column $\overline{v_i}$ is the expected utility of an agent computed with the Markov chain, the fourth column w_i is the share of the value of the optimal CS and the last column v_i is payoff of the agents from our protocol. Note that in the example, the value allocated to the agents from our protocol is much larger than the expected value from traversing the Markov chain.

agent	avg	$\overline{v_i}$	w_i	v_i
0	0.50	0.61	0.17	0.96
1	0.49	0.63	0.17	0.99
2	0.50	0.60	0.16	0.93
3	0.51	0.64	0.18	1.00
4	0.56	0.54	0.15	0.85
5	0.50	0.58	0.16	0.90

Table 2: Agents utilities for a random valuation function

4.4 Discussion on the payoff distribution

Our protocol uses global properties of the valuation function and shares the utility of the optimal CS, $s^\star$, in a fair manner. The distribution of the valuation of $s^\star$, however, is not according to the actual coalitions present in $s^\star$. In other words, given the payoff function v_i, it is possible that, for each coalition $\mathcal{C} \in s^\star$, $\sum_{i \in \mathcal{C}} u_i \neq \sum_{i \in \mathcal{C}} v_i(s^\star)$.

This is different from the traditional assumption in game theory where agents share the value of their coalition. For some agents i, $v_i(s^\star) > u_i$, which may not appear fair. What we propose to the agents is to sign a binding contract to form $s^\star$ and receive u_i as a payoff. If one agent does not want to sign the contract, the agents can form a random CS and try to find a stable CS[3]. From Proposition 2, we see that the expected utility from such a process is at most as good as the value proposed by the protocol and hence the agents have an incentive to accept the guaranteed value while saving on the "cost" of continual change. Hence, on one hand, we want the entire population of agents to cooperate and work together, which has a flavor of using the grand coalition. On the other hand, we want to use the synergy between the agents, and thus form a CS that maximizes social welfare. The reward the agent obtain is designed to be fair for all agents and reflects the performance of the agents over all CSs.

To compute the expected utility of an agent, we have assumed that the coalition formation process starts in a CS picked randomly from a uniform distribution. Of course, some

[3]Note that some agents may benefit from starting the coalition formation process in $s^\star$, hence, if some agents deviate, other agents should force the restart the coalition formation process from a random CS

probability distribution for the initial CSs will benefit some agents in detriment of others. We believe that the probability distribution of the initial CS is part of the definition of the coalition formation problem, and agents do not have any control over it. It is from the entire definition of the coalition formation problem that we compute the expected utility, which we use as a measure of the strength of an agent. If it the distribution is not uniform, the probability to reach the strongly connected components will be different (some components may not be reachable). In addition, the search of the CS that maximizes social welfare should be performed on the subset of CSs that are reachable from the set of possible initial CSs. Minor modification of our computations are needed to address these changes.

5 Conclusion, current and future work

Myopic rational agents who receive a private payoff that depends on the CS may never reach an agreement on the CS to be formed. It may be possible that for each CS, at least one agent has an incentive to change coalition. We designed a protocol that computes a payoff distribution so that agents are guaranteed to have at least the expected utility from a process where each agent would change coalition to maximize its immediate reward. The protocol assumes that 1) the valuation function provides a payoff for each individual agent given a CS and 2) the agents are myopically rational. The payoff function we propose is based on the value of a social welfare maximizing CS and on the expected utility of the agents if they try to change coalitions to maximize their immediate reward. Following our protocol, the agents form the optimal CS, which makes the multiagent system efficient from the viewpoint of a system designer. The valuation of the optimal CS is shared proportionally to the expected utility of the agents. We argue that this is a fair distribution as the payoff obtained by an agent reflects the behavior of the agents over the entire space of CSs, i.e., it is a global property of the valuation function. When the agents follow our protocol, they are guaranteed to have a payoff which is at least their expected value if all agents try to maximize their immediate reward.

The drawback of our approach is its computational cost: the agents needs to build a Markov chain where the number of states is equal to the number of the CSs, which is exponential in the number of agents. Although the corresponding transition matrix is sparse, this method may not be suitable for large number of agents (10 and more). The agents can approximate the expected value by simulating the Markov chain. In that case, they only need to be able to evaluate the best coalitional move from a given CS.

Because of the computational cost, we are studying algorithms to approximate the computation of the Markov chain. By sampling the chain, we can obtain rapidly good estimate of the expected utility of the agents. Another current line of research is the design of protocols and the issue of revealing the valuation function. In the general case, agents have to reveal their valuation, and protocol as [3] can help us ensure that no agent can take advantage of knowledge asymmetry. When agents are sharing a niche, e.g., when the valuation function represent a share attributed to each agent, the agents only need to reveal a preference order over the CSs and no agent has incentive to lie unilaterally.

References

[1] J. Alcalde and A. Romero-Medina. Coalition formation and stability. *Social Choice and Welfare*, 27(2):365–375, 2006.

[2] S. Banerjee, H. Konishi, and T. Sönmez. Core in a simple coalition formation game. *Social Choice and Welfare*, 18(1):135–153, January 2001.

[3] B. Blankenburg, R. K. Dash, S. D. Ramchurn, M. Klusch, and N. R. Jennings. Trusted kernel-based coalition formation. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 989–996, New York, NY, USA, 2005. ACM Press.

[4] A. Bogomolnaiaa and M. O. Jackson. The stability of hedonic coalition structures. *Games and Economic Behavior*, 38(2):201–230, February 2002.

[5] M. Davis and M. Maschler. The kernel of a cooperative game. *Naval Research Logistics Quarterly*, 12, 1965.

[6] T. Dieckmann and U. Schwalbe. Dynamic coalition formation and the core. *Journal of Economic Behavior & Organization*, 49(3):363–380, November 2002.

[7] J. H. Drèze and J. Greenberg. Hedonic coalitions: Optimality and stability. *Econometrica*, 48(4):987–1003, May 1980.

[8] K. K. Fullam, T. B. Klos, G. Muller, J. Sabater, A. Schlosser, Z. Topol, K. S. Barber, J. Rosenschein, L. Vercouter, , and M. Voss. A specification of the agent reputation and trust (ART) testbed: Experimentation and competition for trust in agent societies. In *Proceedings of the Fourth Intenational Joint Conference on Autonomous Agents and Multiagent Systems*, pages 512–518. ACM Press, 2005.

[9] M. Goemans, V. Mirrokni, and A. Vetta. Sink equilibria and convergence. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pages 142–154, Los Alamitos, CA, USA, 2005. IEEE Computer Society.

[10] M. Klusch and A. Gerber. Issues of dynamic coalition formation among rational agents. In *Proceedings of the Second International Conference on Knowledge Systems for Coalition Operations*, pages 91–102, 2002.

[11] T. Rahwan, S. D. Ramchurn, V. D. Dang, and N. R. Jennings. Near-optimal anytime coalition structure generation. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI'07)*, pages 2365–2371, January 2007.

[12] D. Ray and R. Vohra. A theory of endogenous coalition structures. *Games and Economic Behavior*, 26:286–336, 1999.

[13] T. Sandholm and V. R. Lesser. Coalitions among computationally bounded agents. *AI Journal*, 94(1–2):99–137, 1997.

[14] O. Shehory and S. Kraus. Methods for task allocation via agent coalition formation. *Artificial Intelligence*, 101(1-2):165–200, May 1998.

[15] O. Shehory and S. Kraus. Feasible formation of coalitions among autonomous agents in nonsuperadditve environments. *Computational Intelligence*, 15:218–251, 1999.

Stéphane Airiau
ILLC, University of Amsterdam
1018 TV Amsterdam, The Netherlands
Email: `stephane@illc.uva.nl`

Sandip Sen
Computer Science department, University of Tulsa
800 South Tucker dr, Tulsa, OK 74104
Email: `sandip@utulsa.edu`

Computing the Degree of Manipulability in the Case of Multiple Choice

Fuad Aleskerov, Daniel Karabekyan, Remzi Sanver, Vyacheslav Yakuba

Abstract

The problem of the manipulability of known social choice rules in the case of multiple choice is considered. Several concepts of expanded preferences (preferences over the sets of alternatives) are elaborated. As a result of this analysis ordinal and nonordinal methods of preferences expanding are defined. The notions of the degree of manipulability are extended to the case under study. Using the results of theoretical investigation, 22 known social choice rules are studied via computational experiments to reveal their degree of manipulability.

1 Introduction

The problem of manipulation in voting is that the voter can achieve the best social decision for herself by purposely changing her sincere preferences. Theoretical investigations of the manipulation problem were first made in [5] [11]. There, it was shown that if some rather weak conditions hold, any nondictatorial choice rule is manipulable. To which extent social choice rules are manipulable was studied in [1] [7]. However, estimating the degree of manipulability is a very difficult computational problem — to resolve it simplifying assumptions are made. The main and the strongest assumption used is a tie-breaking rule, which allow to consider manipulation problem in the framework of single-valued choice. According to such rule from the set of winning alternatives only one winner is chosen. For example, in [1] [3] in the case of multiple choice the outcome has been chosen with respect to the alphabetical order. This is the most common type of tie-breaking rule because it is simple to implement. But this method also breaks the symmetry between candidates, that can distort the results of computation. The weaker tie-breaking rule was introduced in [9]. According to this rule, a winner is chosen at random in the event of a tie.

Manipulation problem in the case of multiple choice has not been elaborated in detail not only by its computational difficulty, but also because of absence of the common framework allowing to construct preferences over sets of alternatives.

Acknowledgments.

The work of Fuad Aleskerov and Daniel Karabekyan is partially supported by the Scientific Foundation of the Higher School of Economics (grants # 06-04-0052 and # 08-04-0008) and Russian Foundation for Basic Research (grant # 08-01-00039a). The work of Vyacheslav Yakuba is also supported by Russian Foundation for Basic Research (grant # 08-01-00039a).

We express our thanks to all these colleagues and organizations.

2 The framework

We use notations from [1]. There is a finite set A consisting of m alternatives ($m > 2$). Let $\mathcal{A} = 2^A \setminus \{\emptyset\}$ denote a set of all not-empty subsets of the set A. Each agent from a finite set $N = \{1, ..., n\}$, $n > 1$, has preference P_i over alternatives from the set A and expanded preference EP_i over the set $\mathcal{A}$.

Preferences P_i are assumed to be linear orders, i.e., P_i satisfies the following conditions:

- irreflexivity ($\forall x \in A \; x\overline{P}x$),

- transitivity ($\forall x, y, z \in A\ xPy$ and $yPz \Rightarrow xPz$),
- connectedness ($\forall x, y \in A\ x \neq y$ either xPy or yPx).

An ordered n-tuple of preferences P_i is called a profile, $\overrightarrow{P}$. A group decision is made by a social choice rule using $\overrightarrow{P}$ and is considered to be an element of the set $\mathcal{A}$. Let $\mathcal{L}$ denote the set of all linear orders on A . Then the social choice rule can be defined as $F : \mathcal{L}^n \longrightarrow \mathcal{A}$.

Manipulation in the case of multiple choice can be described as follows. Let

$$\overrightarrow{P} = \{P_1, \ldots, P_i, \ldots, P_n\}$$

be the profile of agents' sincere preferences, while

$$\overrightarrow{P}_{-i} = \{P_1, \ldots, P_{i-1}, P_i', P_{i+1}, \ldots, P_n\}$$

is a profile in which all agents but i-th declare their sincere preferences, and P_i' is agent i's deviation from her sincere preference P_i. Let $C(\overrightarrow{P})$, $C(\overrightarrow{P}_{-i})$ denote social choice (a subset of the set $\mathcal{A}$) with respect to profile $\overrightarrow{P}$ and profile $\overrightarrow{P}_{-i}$, correspondingly. Then we say that manipulation takes place if for i-th agent $C(\overrightarrow{P}_{-i}) EP_i C(\overrightarrow{P})$, where EP_i is the expanded preference of i-th agent. In other words, we suppose that outcome when the i-th agent deviates from her true preference is more preferable according to her expanded preference (i.e., according to her preferences over sets) than in the case when she reveals her sincere preference.

3 Basic assumptions for preferences expansion

Let us give some basic conditions of the relationship between preferences over alternatives and expanded preferences over outcomes (sets of alternatives).

First condition was introduced in [6] and is also known as Kelly's Dominance axiom. Here we will use the stronger version of Kelly's axiom introduced in [8]

Kelly's Dominance axiom (strong). $\forall i \in N$ and $\forall \overrightarrow{P}, \overrightarrow{P'} \in \mathcal{L}^n$, if

$$\begin{array}{l} \Big(\forall x \in C(\overrightarrow{P})\ and\ \forall y \in C(\overrightarrow{P'})\ \Rightarrow xP_i y\ or\ x = y\Big)\ and \\ \qquad \Big(\exists z \in C(\overrightarrow{P})\ and\ \exists w \in C(\overrightarrow{P'}) \Rightarrow zP_i w\Big) \end{array},$$

then $C(\overrightarrow{P})\,(EP\,(P_i))\,C(\overrightarrow{P'})$.

We should notice, that this assumption allows us to compare social choices which have at most one alternative in the intersection.

Example. Let $x_1 P_i x_6 P_i x_7 P_i x_9$. Using this condition, we can say that $\{x_1, x_6\}\, EP_i\, \{x_6, x_9\}$, but we cannot compare sets $\{x_1, x_6, x_7\}$ and $\{x_6, x_7, x_9\}$.

In other words, if two outcomes are different only by one alternative, the set which has more preferable alternative must be more preferable than another set.

Gärdenfors principle. $\forall i \in N$, $\forall \overrightarrow{P} \in \mathcal{L}^n$ and $\forall y \in A / C(\overrightarrow{P})$

1) $\Big(C(\overrightarrow{P})\Big)\, EP_i\, \Big(C(\overrightarrow{P}) \cup \{y\}\Big)$ whenever $\forall x \in C(\overrightarrow{P}) : xP_i y$

2) $\Big(C(\overrightarrow{P}) \cup \{y\}\Big)\, EP_i\, \Big(C(\overrightarrow{P})\Big)$ whenever $\forall x \in C(\overrightarrow{P}) : yP_i x$

This condition is also known as Gärdenfors principle defined in [4]. It can be explained in the following way. If we add to some set an alternative which is more (respectively,

less) preferable than every alternative in the chosen set, new outcome should be more (respectively, less) preferable than the old one.

In the literature, for example in [2], another conditions can be found. But almost all of them do not allow us to compare every possible sets of alternatives. For example, for lexicographic preferences ($x_1 P_i x_2 P_i \dots P_i x_{m-1} P_i x_m$) we can not compare sets $\{x_1, x_6, x_7\}$ and $\{x_2, x_4\}$ or $\{x_1, x_{100}\}$ and $\{x_{99}, x_{101}\}$. Thus, we should define algorithms of preferences expanding which satisfy conditions mentioned above and allow us to compare all the sets of alternatives.

4 Preference expanding methods

4.1 Lexicographic methods

4.1.1 Leximin

This algorithm of preferences expansion is introduced in [8] and is based on the well-known maximin behaviour approach. Here we will use it in the form given in [10]. This method is based on comparison of the worst alternatives of any two sets. If the worst alternatives are the same, then we should compare second-worst altenatives and so on. If this is impossible, that is, when one social choice is a subset of another social choice, then the greater set is preferred to the lesser one.

Let us describe leximin method of preferences expanding. From preferences $P_i \in \mathcal{L}$ we can recieve leximin expanded preferences EP_i by the following algorithm.

Two social choices $X, Y \in \mathcal{A}$ are compared:

1. If $|X| = |Y| = k$, where $k \in \{1, \dots, m-1\}$, then sort alternatives from each social choice from the most preferred to the least one, that is: $X = \{x_1, \dots, x_k\}$ and $Y = \{y_1, \dots, y_k\}$, where $x_j P_i x_{j+1}$ and $y_j P_i y_{j+1}$ $\forall j \in \{1, \dots, k-1\}$. Then $X\ EP_i\ Y$ if and only if $x_h P_i y_h$ for the greatest $h \in \{1, \dots, k\}$ for which $x_h \neq y_h$.

2. If $|X| \neq |Y|$, then sort alternatives from each social choice from the least preferred to the most one, that is: $X = \{x_1, \dots, x_{|X|}\}$ and $Y = \{y_1, \dots, y_{|Y|}\}$, where $x_{j+1} P_i x_j$ $\forall j \in \{1, \dots, |X|-1\}$ and $y_{j+1} P_i y_j$ $\forall j \in \{1, \dots, |Y|-1\}$. There can be two cases:

   (a) $x_h = y_h$ $\forall h \in \{1, \dots, \min\{|X|, |Y|\}\}$. That is, one social choice is a subset of another social choice. Then, it was already mentioned above, the greater set is preferred to the lesser one, that is, $X\ EP_i\ Y$ if and only if $|X| > |Y|$.

   (b) $\exists h \in \{1, \dots, \min\{|X|, |Y|\}\}$ for which $x_h \neq y_h$. Then $X\ EP_i\ Y$ if and only if $x_h P_i y_h$ for the least $h \in \{1, \dots, \min\{|X|, |Y|\}\}$, for which $x_h \neq y_h$.

For example, for three alternatives and preferences $a P_i b P_i c$ over them, leximin expanded preferences EP_i will be

$$\{a\}\ EP_i\ \{a,b\}\ EP_i\ \{b\}\ EP_i\ \{a,c\}\ EP_i\ \{a,b,c\}\ EP_i\ \{b,c\}\ EP_i\ \{c\}$$

4.1.2 Leximax

This preferences expanding method is similar to the leximin one, but in this case the best of any two social choices are compared. If the best alternatives are the same, then we should compare second-best altenatives and so on. If this is impossible, that is, when one social choice is a subset of another social choice, then the lesser set is preferred to the greater one.

For example, for three alternatives and preferences $a P_i b P_i c$ over them, leximax expanded preferences EP_i will be

$$\{a\}\, EP_i\, \{a,b\}\, EP_i\, \{a,b,c\}\, EP_i\, \{a,c\}\, EP_i\, \{b\}\, EP_i\, \{b,c\}\, EP_i\, \{c\}$$

4.2 Probabilistic methods

These methods of preferences expanding in contrast to lexicographic methods suggest that for voter not only the presence of the alternative in a social choice is important, but the probability that this alternative would be the final outcome is important as well. Here two algorithms are considered: an ordering is constructed based on the probability of the best alternative and an ordering is constructed based on the probability of the worst alternative.

4.2.1 Ordering based on the probability of the best alternative

This preference expanding algorithm is based on the element-wise comparison of two social choices. If the best alternatives of two sets are the same, then the set, in which the probability that this alternative would be the final outcome is higher, is more preferable. In fact, it will be the lesser set. If the best alternatives are the same and have equal probability to be the final outcome, then next alternatives are compared in the same way.

Example. In the set $\{a,b,c\}$ probability that alternative a would be the final outcome equals $\frac{1}{3}$ (we assume that each alternative of the winning set has equal probability to be chosen as final outcome). In the set $\{a,c\}$ this probability equals $\frac{1}{2}$. In other words, there will be $\{a,c\}\, EP_i\, \{a,b,c\}$ by expanded preferences based on the probability of the best alternative algorithm.

For example, for three alternatives and preferences aP_ibP_ic over them, expanded preferences EP_i based on the probability of the best alternative will be:

$$\{a\}\, EP_i\, \{a,b\}\, EP_i\, \{a,c\}\, EP_i\, \{a,b,c\}\, EP_i\, \{b\}\, EP_i\, \{b,c\}\, EP_i\, \{c\}$$

4.2.2 Ordering based on the probability of the worst alternative

This preferences expanding method is similar to the previous, but in this case the probability of the worst alternative is consider. The set in which this probability is higher is less preferable.

For example, for three alternatives and preferences aP_ibP_ic over them, expanded preferences EP_i based on the probability of the worst alternative will be:

$$\{a\}\, EP_i\, \{a,b\}\, EP_i\, \{b\}\, EP_i\, \{a,b,c\}\, EP_i\, \{a,c\}\, EP_i\, \{b,c\}\, EP_i\, \{c\}$$

4.3 Ordinal methods

This approach is based on the assumption of expected utility maximization introduced by von Neuman and Morgenstern. Here we will use a particular case of this assumption.

1. First of all, we will assign utility of each alternative for its place in preferences. In fact, we will rank the alternatives - the best one will receive the rank m, the next one the rank $m-1$, and so on. The worst alternative has the rank of 1.

2. We assume that each alternative has equal probability to be chosen as the final outcome. It means that utility of the set of alternatives is equal to the average utility value of all alternatives within this set.

In fact, even these assumptions do not allow us to compare all social choices when $m > 2$. For example, for three alternatives and preferences aP_ibP_ic over them, there are sets $\{a,b,c\}$, $\{a,c\}$, $\{b\}$, which have equal utility of 2 according to this approach. So, we need to consider additional assumptions.

4.3.1 Lexicographic expansions

These methods suggest the use of lexicographic approach to the sets which are uncompared by ordinal method itself. Note that new expanded preferences may differ from lexicographic preferences.

4.3.2 Probabilistic expansions

This methods suggest the use of probabilistic approach to the sets which are uncompared by ordinal method itself. For example, for four alternatives and preferences $aP_ibP_icP_id$ over them, expanded preferences EP_i based on ordinal method with the probability of the worst alternative approach are (the groups of the sets for which expansion is used are underlined):

$$\{a\}\, EP_i\, \{a,b\}\, EP_i \underline{\{b\}\, EP_i\, \{a,b,c\}\, EP_i\, \{a,c\}} EP_i$$
$$EP_i\, \{a,b,d\}\, EP_i \underline{\{b,c\}\, EP_i\, \{a,b,c,d\}\, EP_i\, \{a,d\}} EP_i$$
$$EP_i\, \{a,c,d\}\, EP_i \underline{\{c\}\, EP_i\, \{b,c,d\}\, EP_i\, \{b,d\}} EP_i\, \{c,d\}\, EP_i\, \{d\}$$

4.3.3 Attitude to risk expansions

These methods are based on attitude to risk approach. In the case when the expected utility of several sets is equal, the risk-averse voter will prefer the set with the lowest variance and risk-lover voter will prefer the set with the highest variance. For up to 6 alternatives expanded preferences based on ordinal method with this expansions coincide with expanded preferences based on ordinal method with probabilistic expansion. If the number of alternatives is greater than 7 the coincidence does not hold.

Example. Let us consider lexicographic preferences $x_1P_ix_2P_i \dots P_ix_n$, where $n \geq 7$. There are sets $\{x_1, x_5, x_6\}$ and $\{x_2, x_3, x_7\}$ which have the equal rank and the equal variance. So, these sets are uncompared by ordinal method with attitude to risk expansions.

For three alternatives these methods yield the same results as probabilistic methods, but for four alternatives this fact does not hold. For example, for four alternatives and the preferences $aP_ibP_icP_id$ over them, expanded preferences EP_i based on ordinal method with risk-lover expansion are (the groups of the sets for which expansion is used are underlined):

$$\{a\}\, EP_i\, \{a,b\}\, EP_i \underline{\{a,c\}\, EP_i\, \{a,b,c\}\, EP_i\, \{b\}} EP_i$$
$$EP_i\, \{a,b,d\}\, EP_i \underline{\{a,d\}\, EP_i\, \{a,b,c,d\}\, EP_i\, \{b,c\}} EP_i$$
$$EP_i\, \{a,c,d\}\, EP_i \underline{\{b,d\}\, EP_i\, \{b,c,d\}\, EP_i\, \{c\}} EP_i\, \{c,d\}\, EP_i\, \{d\}$$

4.3.4 Cardinality expansions

This approach is based on comparison of the cardinality of sets which are uncomapred by ordinal method itself. We assume, that when expected utility of several sets is equal, then for voter a cardinality is important. There are two methods: one assumes that the greater set is preferred to smaller one in case of the same rank, and the other assume that the smaller set is preferred to the greater one. Note that these assumptions are rather non-binding. It allows us to compare all sets only when there are three alternatives. However, even in this case this method do not give different results. For example, for three alternatives and preferences aP_ibP_ic over them, expanded preferences EP_i based on ordinal method with greater set approach yield the same result as leximax method and ordinal method with leximax expansion:

$$\{a\}\, EP_i\, \{a,b\}\, EP_i\, \{a,b,c\}\, EP_i\, \{a,c\}\, EP_i\, \{b\}\, EP_i\, \{b,c\}\, EP_i\, \{c\}$$

For more than four alternatives cardinality expansion itself don't allows to compare all sets of alternatives. So, additional expansions mentioned above should be added in this case.

5 Indices of manipulability

Number of alternatives being m, the total number of possible linear orders is obviously equal to $m!$, and total number of profiles with n agents is equal to $(m!)^n$. In [7] to measure a degree of manipulability of social choice rules the following index was introduced (we call it Kelly's index and denote as K) :

$$K = \frac{d_0}{(m!)^n},$$

where d_0 is the number of profiles in which manipulation takes place[1].

In [1] index of freedom of manipulation is introduced. We also introduce two similar indices: the degree of nonsensitivity to preference change and probability of getting worse. Let us note that for an agent there are $(m! - 1)$ linear orders to use instead of her sincere preference. Denote as κ_i^+ $(i = 1, ..., n;\ 0 \leq \kappa_i^+ \leq m! - 1)$ the number of orderings in which voter is better off, κ_i^0 - the number of orderings when the result of voting remain the same and κ_i^- -the number of orderings in which voter is worse off. It is obvious that $\kappa_i^+ + \kappa_i^0 + \kappa_i^- = (m! - 1)$. Dividing each κ_i to $(m! - 1)$ one can find the share of each type of orderings for an agent i in this profile. Summing up each share over all agents and dividing it to n one can find the average share in the given profile. Summing the share over all profiles and dividing this sum to $(m!)^n$ we obtain three indices

$$I_1 = \frac{\sum_{j=1}^{(m!)^n} \sum_{i=1}^{n} \kappa_i}{(m!)^n \cdot n \cdot (m!-1)}$$

where κ_i is κ_i^+, κ_i^0 or κ_i^-. It is obvious that $I_1^+ + I_1^0 + I_1^- = 1$.

These indices K and I_1 (as well as index J) measure the degree of manipulability in terms of the share of manipulable profiles or the share of orderings using which an agent can manipulate.

The following two indices show the *efficiency* of manipulation, i.e., to which extent an agent can be better off via manipulating her sincere ordering. Let under a profile $\overrightarrow{P}$ social decision be the set $C\left(\overrightarrow{P}\right)$ which stands at k-th place from the top in the expanded preferences of i-th agent. Let after her manipulation the social decision be a set $C\left(\overrightarrow{P'}\right)$ which stands in the expanded preferences of the i-th agent at j-th place from the top, and let $j < k$. Then $\theta = j - k$ shows how is the i-th agent better off. Let us sum up θ for all advantageous orderings κ_i^+ (defined above), and let us divide the obtained value to κ_i^+. Denote this index through Z_i, which shows an average "benefit" (in terms of places) of manipulation of the agent i gained via manipulaiton κ_i^+ orderings from $(m! - 1)$. Summing up this index over all agents and over all profiles, we obtain the index under study

$$I_2 = \frac{\sum_{j=1}^{(m!)^n} \sum_{i=1}^{n} Z_i}{(m!)^n \cdot n}$$

The next criterion I_3 is a modification of I_2. Instead of evaluating the "average" benefit Z_i for i-th agent, we evaluate the value

$$Z_i^{\max} = \max(Z_1, ..., Z_{\kappa_i}).$$

In other words, the value $Z_i^{\max}$ show the maximal benefit which can be obtained by agent i. Summing up this index over all agents and over all profiles, we obtain our next index under study

[1] In [1] an extended version of Kelly's index was introduced. Denote by λ_k the number of profiles in which exactly k voters can manipulate. Construct index $J_k = \frac{\lambda_k}{(m!)^n}$ which shows the share of profiles in which exactly k voters can manipulate. Obviously, $K = J_1 + J_2 + ... + J_n$. Then one can consider the vectorial index $J = (J_1, J_2, ..., J_n)$.

$$I_3 = \frac{\sum_{j=1}^{(m!)^n} \sum_{i=1}^{n} Z_i^{\max}}{(m!)^n \cdot n}$$

The indices K, I_1, I_2, I_3, J have been calculated for each of the rules introduced in the next section. The indices I_1, I_2 and I_3 were introduced in [1].

6 Social Choice Rules

The calculation of indices is performed for up to m = 5 alternatives for 22 social choice rules. In this work the results only for 5 rules will be given.

1. Plurality Rule

Choose alternatives, that have been admitted to be the best by the maximum number of agents, i.e.

$$a \in C(\overrightarrow{P}) \Longleftrightarrow [\forall x \in A \quad n^+(a, \overrightarrow{P}) \geq n^+(x, \overrightarrow{P})],$$

where $n^+(a, \overrightarrow{P}) = card\{i \in N \mid \forall y \in A \;\; aP_iy\}$

2. Approval Voting.

Let us define

$$n^+(a, \overrightarrow{P}, q) = card\{i \in N \mid card\{D_i(a)\} \leq q - 1\},$$

i.e., $n^+(a, \overrightarrow{P}, q)$ means the number of agents for which a is placed on q'th place in their orderings. Thus, if $q = 1$, then a is the first best alternative for i-th voter; if $q = 2$, then a is either first best or second best option, etc. The integer q can be called as degree of procedure.

Now we can define Approval Voting Procedure with degree q

$$a \in C(\overrightarrow{P}) \Leftrightarrow [\forall x \in A \; n^+(a, \overrightarrow{P}, q) \geq n^+(x, \overrightarrow{P}, q)],$$

i.e., the alternatives are chosen that have been admitted to be between q best by the maximum number of agents.

It can be easily seen that Approval Voting Procedure is a direct generalization of Plurality Rule; for the latter $q = 1$.

3. Borda's Rule.

Put to each $x \in A$ into correspondence a number $r_i(x, \overrightarrow{P})$ which is equal to the cardinality of the lower contour set of x in $P_i \in \overrightarrow{P}$, i.e. $r_i(x, \overrightarrow{P}) = |L_i(x)| = |\{b \in A : xP_ib\}|$. The sum of that numbers over all $i \in N$ is called Borda's count for alternative x,

$$r(a, \overrightarrow{P}) = \sum_{i=1}^{n} r_i(a, P_i).$$

Alternative with maximum Borda's count is chosen., i.e.

$$a \in C(\overrightarrow{P}) \Longleftrightarrow [\forall b \in A, \quad r(a, \overrightarrow{P}) \geq r(b, \overrightarrow{P})].$$

4. Black's Procedure.

If Condorset winner exists, it is to be chosen. Otherwise, Borda's Rule is applied.

5. Threshold rule.

Let $v_1(x)$ be the number of agents for which the alternative x is the worst in their ordering, $v_2(x)$ is the number of agents placing the x second worst, and so on, $v_m(x)$

the number of agents considering the alternative x the best. Then we order the alternatives lexicographically. The alternative x is said to V-dominate the alternative y if: $v_1(x) < v_1(y)$ or, if there exists k not more than m, s.t. $v_i(x) = v_i(y)$, $i = 1, ..., k-1$, and $v_k(x) < v_k(y)$. In other words, first, the number of worst places are compared if these numbers are equal then the number of second worst places are compared and so on. The alternatives which are not dominated by other alternatives via V are chosen.

7 Computation scheme

The calculation of indices is performed for up to m = 5 alternatives. For small number of voters n all possible profiles are checked for manipulability and respective indices are evaluated. For greater number of voters the statistical scheme is used.

For each profile under consideration in both exhaustive and statistical schemes, all m!-1 manipulating orderings for each voter are generated and the respective choice sets of manipulating profiles are compared with the choice of the original profile, using all introduced methods of the preference extensions.

8 Results

In the case of 3 alternatives there are 4 different types of extended preferences. For example, if preferences over are aP_ibP_ic then the extended preferences are as follows:

1. Leximin method, Ordinal method with leximin or greater set extensions.

$$\{a\}\, EP_i\, \{a,b\}\, EP_i\, \{b\}\, EP_i\, \{a,c\}\, EP_i\, \{a,b,c\}\, EP_i\, \{b,c\}\, EP_i\, \{c\}$$

2. Leximax method, Ordinal method with leximax or lesser set extensions.

$$\{a\}\, EP_i\, \{a,b\}\, EP_i\, \{a,b,c\}\, EP_i\, \{a,c\}\, EP_i\, \{b\}\, EP_i\, \{b,c\}\, EP_i\, \{c\}$$

3. Probabilistic method based on the probability of the worst alternative, Ordinal method with risk-averse extension.

$$\{a\}\, EP_i\, \{a,b\}\, EP_i\, \{b\}\, EP_i\, \{a,b,c\}\, EP_i\, \{a,c\}\, EP_i\, \{b,c\}\, EP_i\, \{c\}$$

4. Probabilistic method based on the probability of the best alternative, Ordinal method with risk-lover extension.

$$\{a\}\, EP_i\, \{a,b\}\, EP_i\, \{a,c\}\, EP_i\, \{a,b,c\}\, EP_i\, \{b\}\, EP_i\, \{b,c\}\, EP_i\, \{c\}$$

The results of Kelly's index calculation for 3 and 4 voters are presented in the tables 1 and 2 correspondingly. In the brackets near the name of the rule the results from [1] are given. One can see that in most cases, especially in the case of 4 voters, the degree of manipulability in the case of single-valued choice is underestimated. We also can state that for almost all rules Method 1 and Method 3 have the same Kelly's index as well as Methods 2 and 4.

In Figures 1 and 2 the results of calculation for the larger number of voters are given. Kelly's index is shown on the Y-axis and the logarithm of the number of voters is shown on the X-axis. The calculation was made for each number of voters from 3 to 25 and then for 29, 30, 39, 40 and so on up to 100. That explains changes at the figures when number of voters is more than 25.

We can make several conclusions from these figures.

	Method 1	Method 2	Method 3	Method 4
Plurality (0,1667)	0,2222	0	0,2222	0
Approval q=2	0,1111	0,6111	0,1111	0,6111
Borda (0,2361)	0,3056	0,4167	0,3056	0,4167
Black (0,1111)	0,0556	0,1667	0,0556	0,1667
Threshold	0,3056	0,4167	0,3056	0,4167

Table 1: The case of 3 alternatives and 3 voters

	Method 1	Method 2	Method 3	Method 4
Plurality (0,1852)	0,3333	0,3333	0,3333	0,3333
Approval q=2	0,2963	0,2963	0,2963	0,2963
Borda (0,3102)	0,3611	0,4028	0,3611	0,4028
Black (0,1435)	0,2361	0,2778	0,2778	0,2361
Threshold	0,4028	0,4028	0,4028	0,4028

Table 2: The case of 3 alternatives and 4 voters

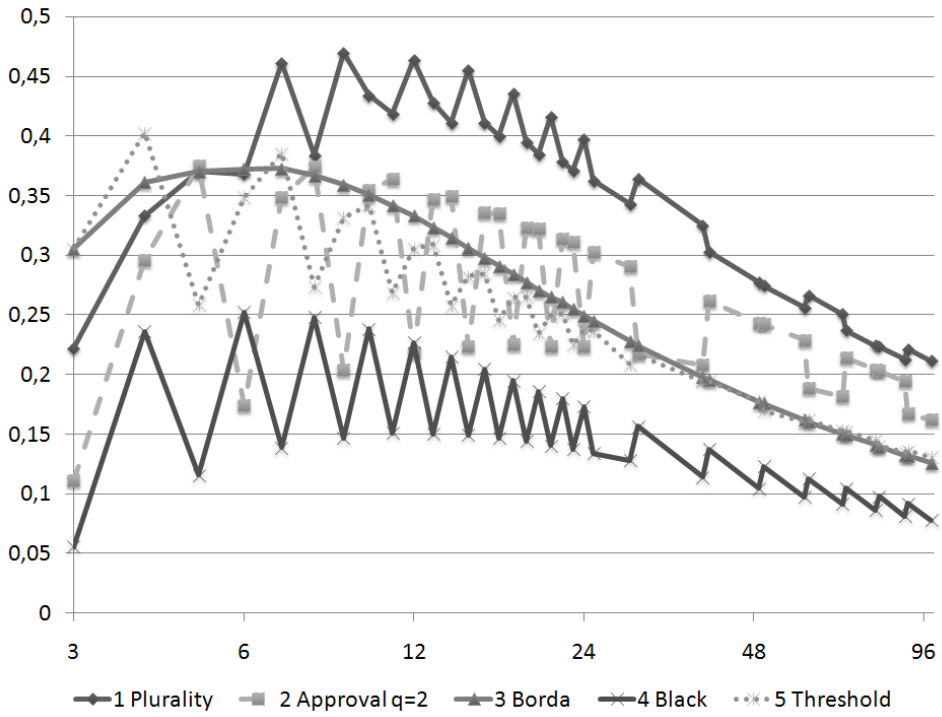


Figure 1: Kelly's index for Leximin extension method.

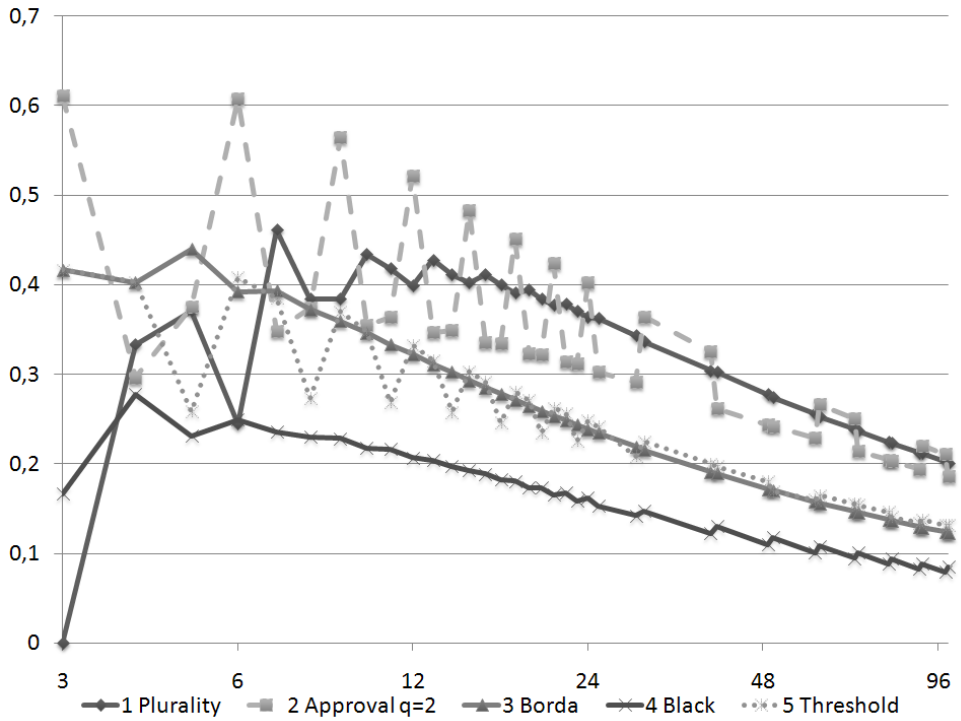


Figure 2: Kelly's index for Leximax extension method.

1. The answer on the question which rule is less manipulable depends on the method of preferences extension. For example, for the number of voters divisible by 3 Approval voting rule is less manipulable than Plurality rule for Method 1 and is more manipulable for Method 2.

2. If the number of voters is small Threshold rule is less manipulable than Borda rule. But when the number of voters is high enough Borda rule is better in Kelly's sense. The exact minimum number of voters needed depends on the method used.

3. Black's Procedure is least manipulable almost for any number of voters and for any method.

4. Kelly's index for Black's Procedure and Method 1 depends on even or odd number of the voters considered. At the same time for rules such as Plurality, Approval voting and Threshold, there is a cycle length of m. In this case there is a cycle length of 3. The dependence from the number of alternatives is explained by differences in number and cardinality of ties produced by rules. For example, the set $\{a, b, c\}$ can appear as the result of plurality voting only in the case when the number of voters is divisible by the number of alternatives.

In Figure 3 the results of calculation of I_1 index for 3 alternatives, 3 voters and Method 1 are given. The left part of each row is the degree of freedom of manipulation. The right part is the probability of getting worse. The middle part is the degree of nonsensitivity to preference change. In Figure 4 the results of calculation of I_1 index for 3 alternatives, 100 voters and Method 1 are given. We can see that the bigger the number of voters is, the less is freedom of manipulation as well as the probability of getting worse.

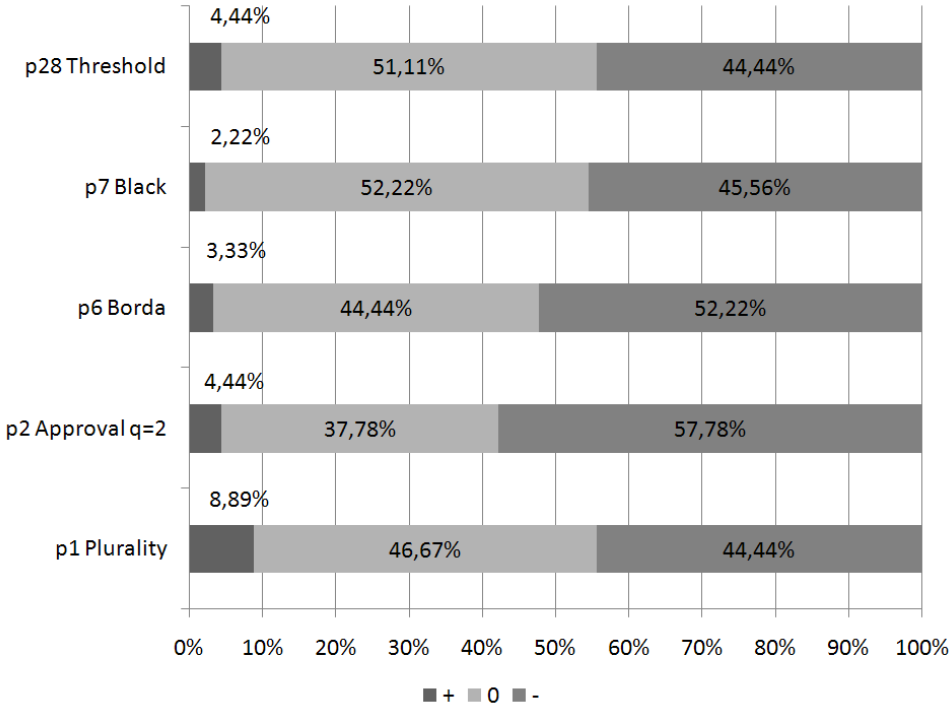


Figure 3: I_1 for Leximin extension method and 3 alternatives.

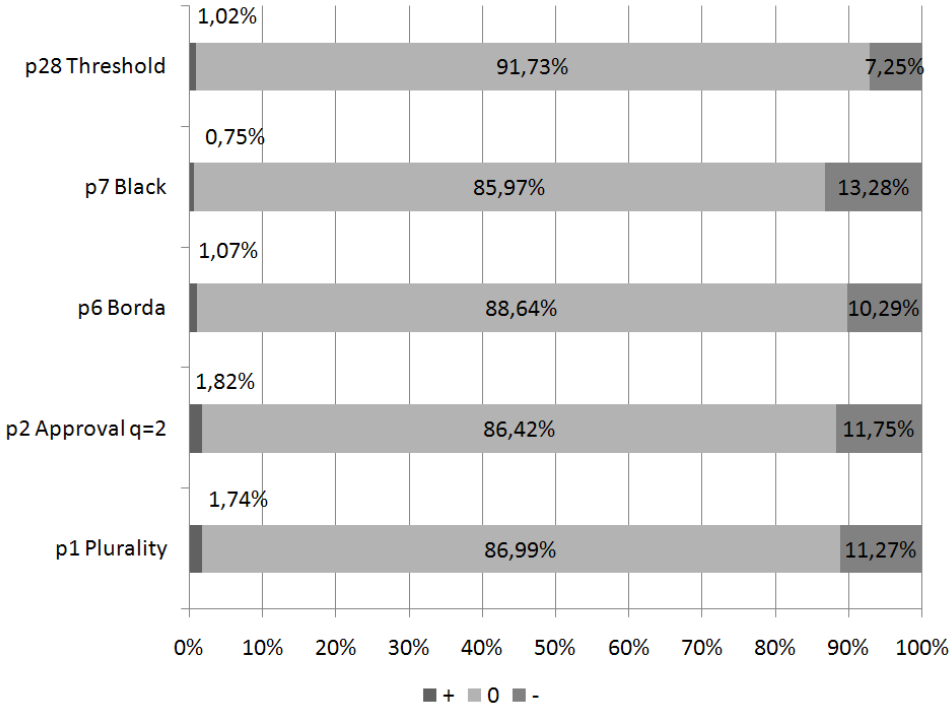


Figure 4: I_1 for Leximin extension method and 100 alternatives.

References

[1] Aleskerov F, Kurbanov E (1999) "Degree of manipulability of social choice procedures", in: Alkan et al. (eds.) Current Trends in Economics. Springer, Berlin Heidelberg New York

[2] Barbera S (1977) "The manipulability of social choice mechanisms that do not leave too much to chance", Econometrica Volume 45: 1572-1588

[3] Favardin P, Lepelley D (2006) "Some further results on the manipulability of social choice rules", Social Choice and Welfare Volume 26: 485-509

[4] Gärdenfors P (1976) "Manipulation of social choice functions", Journal of Economic Theory Volume 13: 217-228

[5] Gibbard A (1973) "Manipulation of voting schemes", Econometrica Volume 41: 587-601

[6] Kelly J (1977) "Strategy-proofness and social choice functions without single valuedness" Econometrica Volume 45: 439–4

[7] Kelly J (1993) "Almost all social choice rules are highly manipulable, but few aren't", Social Choice and Welfare Volume 10: 161–175

[8] Pattanaik P (1978) "Strategy and group choice", North-Holland, Amsterdam

[9] Pritchard G, Wilson M (2005) "Exact results on manipulability of positional voting rules", CDMTCS Research Report Series

[10] Sanver R., Ozyurt S. (submitted) "A general impossibility result on strategy-proof social choice hyperfunctions", Submitted to Games and Economic Behavior

[11] Satterthwaite M (1975) "Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions", Journal of Economic Theory Volume 10: 187-217

[12] Zwicker W (2005) "Manipulability, Decisiveness, and Responsiveness", research proposal

Fuad Aleskerov
Department of Economics, State University - Higher School of Economics
Pokrovsky blvd, 11, office Zh 428, Moscow, 109028, Russia
Email: `alesk@hse.ru`

Daniel Karabekyan
Department of Economics, State University - Higher School of Economics
Pokrovsky blvd, 11, office Zh 428, Moscow, 109028, Russia
Email: `danyakar@gmail.com`

M. Remzi Sanver
Department of Economics, Istanbul Bilgi University
Inonu Cad. No. 28, Kustepe, Istanbul, 80310, Turkey
Email: `sanver@superonline.com`

Vyacheslav Yakuba
Institute of Control Science
65, Profsoyuznaya str, office 324, Moscow, 117997, Russia
Email: `yakuba@ipu.ru`

On the Complexity of Rationalizing Behavior

Jose Apesteguia and Miguel A. Ballester

Abstract

We study the complexity of rationalizing choice behavior. We do so by analyzing two polar cases, and a number of intermediate ones. In our most structured case, that is where choice behavior is defined in universal choice domains and satisfies the "weak axiom of revealed preference," finding the complete preorder rationalizing choice behavior is a simple matter. In the polar case, where no restriction whatsoever is imposed, either on choice behavior or on choice domain, finding the complete preorders that rationalize behavior turns out to be intractable. We show that the task of finding the rationalizing complete preorders is equivalent to a graph problem. This allows the search for existing algorithms in the graph theory literature, for the rationalization of choice.

1 Introduction

The theory of individual decision making is typically formulated on the basis of two different approaches. A revealed preference approach that directly studies individual choice behavior through choice correspondences. Or a preference relation approach, where tastes are summarized through binary relations. Clearly, the relation between these two distinct formal approaches to individual behavior is a fundamental question in economics. It is said that a collection of preference relations rationalizes choice behavior, whenever both approaches identify the same alternatives for every possible choice problem that may arise.[1] This problem has attracted a great deal of attention in the literature.

Drawing on the tools of theoretical computer science, we study the question of how complex it is to find the preference relations that rationalize choice behavior. That is, given any possible choice rule, how difficult it is to construct the collection of preference relations that summarizes choice behavior. In spite of the fact that the question of rationalization is central to economics, to the best of our knowledge this paper is the first one to study the practical difficulty of rationalization.

We analyze two polar cases and some intermediate ones. In the first place we study what we call the rational procedure. This is the case where the choice correspondence satisfies the well-known consistency property known as the "weak axiom of revealed preference" (WARP). The classic result in this case is that a choice correspondence defined on the universal choice domain (i.e., choice is defined on all possible choice problems) satisfying WARP is rationalized by a unique complete preorder. We show that finding the complete preorder in such a case is a simple matter, as there are algorithms that can easily construct it.

We then turn the analysis to the polar case where no restriction whatsoever is imposed, neither on choice behavior nor on the choice domain. In other terms, we do not impose any consistency property on choice behavior; neither WARP, nor any other property. With regard to the choice domain, we consider arbitrary collections of choice problems. Then, drawing from a seminal paper by Kalai, Rubinstein, and Spiegler (2002; hereafter KRS), in order to rationalize behavior we use a collection (a book) of complete preorder relations (rationales), such that for every choice problem A in the domain of feasible choice problems, the choice $c(A)$ is maximal in A for some complete preorder. Clearly, there are multiple books that rationalize a given choice rule. KRS naturally propose to focus on those books

[1] In the next section we will make precise the terminology we use in this introduction.

that use the minimal number of preference relations. In this case, we show that, contrary to the previous one, finding a minimal book is a difficult computational problem (see Theorem 3.2). That is, there is little hope for the existence of an algorithm that for every possible choice rule finds a minimal book in a reasonable time frame. Hence, in this extremely unstructured case, the task of finding the collection of preferences that rationalize behavior is in general intractable.

Now, the question arises whether it is the conjunction of (i) unstructured choice behavior and (ii) unrestricted choice domain that leads to the computational hardness of the problem of rationalization. It could be the case that the difficulty in finding the preference relations rationalizing choice comes from the unstructured nature of behavior. It turns out that the answer to this question seems to depend on the single or multi-valued nature of the choice function.

Theorem 3.4 suggests that in the case of single-valued choice functions, the essence of the intractability of rationalization is triggered by the interplay of both unstructured behavior and unrestricted domain. Theorem 3.4 shows that under the universal choice domain, the problem of finding a minimal book is quasi-polynomially bounded. However, we argue that in the case of choice correspondences the practical difficulty of rationalizing behavior may be due to the nature of choice behavior per se.

The challenge is then to understand the driving force of the complexity of rationalization. If we are able to understand the roots of the complexity in rationalizing choice behavior, we may use this to search for specific algorithms that behave well under certain circumstances. We start this challenge by defining two binary relations on the space of choice problems, that capture two fundamental properties on the structural relation of choice problems. By doing so we will be able to draw a connection with a graph theory problem (see Theorem 4.3). This is especially useful since there is a wealth of algorithms for graph problems that may be used to solve the problem of rationalization of certain choice structures.

We then end the analysis by exploring a specific case, inspired by the recent work of Manzini and Mariotti (2007; hereafter MM). MM study the nature of choice functions that can be rationalized by sequentially applying a fixed set of asymmetric binary relations. Among other results, they provide a full characterization for the case when a choice function is sequentially rationalizable by two rationales. Such a choice function is called a Rational Shortlist Methods (RSMs). Using the tools derived for establishing the connection with graph theory, we show that the rationalization of RSMs through multiple rationales turns out to be a computationally tractable problem.

Apart from those papers already mentioned, Salant (2003), Ok (2005), and Xu and Zhou (2007) constitute recent related works. Salant (2003) studies two computational aspects of choice when the "independence of irrelevant alternatives" (IIA) axiom does not necessarily hold: the amount of memory choice behavior requires, and the computational power needed for the computation of choice. He shows that the rational procedure is favored by these considerations. Ok (2005) provides an axiomatic characterization of choice correspondences that satisfy the IIA axiom. Finally, Xu and Zhou (2007) propose the rationalization through extensive games with perfect information, and provide a full characterization of those choice functions that can be so rationalized.

The rest of the paper is organized as follows. Section 2 gives the definitions on choice behavior and binary relations, and makes precise the notion of rationalization we will use throughout the paper. Furthermore, it contains a brief introduction to the theory of NP-completeness. Section 3 contains the complexity results. In section 4 we draw on the connection between the problem of rationalization and the literature on graph theory. Finally, section 5 concludes and relates our work to the economics literature on language and to the bounded rationality literature.

2 Preliminaries

2.1 Choice behavior and preference relations

Let X be a finite set of n objects. We denote by $\mathcal{U}$ the set of all non-empty subsets of X. We also consider the general case of arbitrary domains $\mathcal{D} \subseteq \mathcal{U}$, with $\mathcal{D} \neq \emptyset$. A choice correspondence c on $\mathcal{D}$ assigns to every $A \in \mathcal{D}$ a non-empty set of elements $c(A) \subseteq A$. A single-valued choice function simply assigns to every $A \in \mathcal{D}$, a unique element $c(A) \in A$.[2]

The weak axiom of revealed preference is the classic consistency property:

Weak Axiom of Revealed Preference (WARP): Let $A, B \in \mathcal{D}$ and assume $x, y \in A \cap B$; if $x \in c(A)$ and $y \in c(B)$, then we must also have $x \in c(B)$.

WARP imposes a great deal of structure on choice. This can be best appreciated through its implications on the connection with preference relations. In order to elaborate on the connection between choice behavior and preference relations, which is central to this paper, we first introduce some notation. Denote by $\succeq$ a binary relation on X, $\succeq \subseteq X \times X$. Binary relations $\succ$ and $\sim$ are the asymmetric and symmetric parts of $\succeq$, respectively. Hence, $x \succ y$ if and only if $x \succeq y$ and $\neg(y \succeq x)$, and $x \sim y$ if and only if $x \succeq y$ and $y \succeq x$. We will say that a binary relation is a preorder if it is reflexive and transitive. We will often refer to complete preorders by rationales. We will say that a binary relation is a linear order if it is antisymmetric, transitive, and complete. Finally, we will say that it is a partial order if it is reflexive, antisymmetric, and transitive. For any $A \in \mathcal{D}$, $M(A, \succeq)$ denotes the set of maximal elements in A with respect to $\succeq$, that is, $M(A, \succeq) = \{x \in A : y \succ x \text{ for no } y \in A\}$. Let $c(A) \succeq c(B)$, $A, B \in \mathcal{D}$, denote the case when for every $x \in c(A)$ and $y \in c(B)$, we have $x \succeq y$.

The classic notion of *rationalization* of a choice correspondence deals with the issue of existence of a unique complete preorder relation $\succeq$ that explains choice behavior defined on $\mathcal{U}$.[3] The question is whether there is a $\succeq$ such that $c(A) = M(A, \succeq)$ for every $A \in \mathcal{U}$. A well-known result establishes that a choice correspondence c satisfies WARP if and only if there is a complete preorder $\succeq$ such that $c(A) = M(A, \succeq)$ for every $A \in \mathcal{U}$.[4] This result makes it clear that if c does not satisfy WARP then a broader definition of rationalization is needed.

In the context of single-valued choice functions and universal choice domains $\mathcal{U}$, KRS propose the rationalization of any possible choice function through collections of linear orders. The interest is naturally directed to the minimal number of linear orders that rationalizes choice behavior. Here we use this notion of rationalization. To this end we extend the original definition to include choice correspondences and arbitrary choice domains $\mathcal{D}$. Accordingly, we also substitute linear orders by complete preorders.

Minimal Rationalization by Multiple Rationales (RMR): A K-tuple of complete preorders $(\succeq_k)_{k=1,\ldots,K}$ on X is a rationalization by multiple rationales (RMR) of c if for every $A \in \mathcal{D}$, there is a $k \in \{1, \ldots, K\}$, such that $c(A) = M(A, \succeq_k)$. It is said to be minimal if any other RMR of c has at least K preference relations.

Note that a minimal RMR is an extension of the classic idea of rationalization by one rationale. In fact, if c satisfies WARP, then the minimal RMR is composed of one complete

[2] To avoid tedious duplication of notation, we denote both the multi-valued and the single-valued cases by c. In any case, the context will be specific enough to avoid confusion.

[3] Uzawa (1957), Arrow (1959), Richter (1966) and Sen (1970) were among the first to study aspects of this problem.

[4] Moulin (1985) and Suzumura (1983) provide surveys of the related literature.

preorder. Further, note also that the case of choice correspondences generalizes the case of single-valued choice functions. The classic result in the latter context is that a single-valued choice function c is rationalized by a linear order if and only if c satisfies the property of independence of irrelevant alternatives (IIA):

Independence of Irrelevant Alternatives (IIA): For any $A, B \in \mathcal{D}$, if $y \in A \subseteq B$ and $y \in c(B)$, then $y \in c(A)$.

We highlight an especially important class of choice sets; the collection of c-maximal sets. These are the sets that must be explained in order to rationalize behavior. A set S is c-maximal if any addition to S of elements (consistent with the domain of choice problems $\mathcal{D}$) leads to a change in choice behavior with respect to the original elements of S. Formally,

c-Maximal Sets: A subset $S \in \mathcal{D}$ is said to be c-maximal if for all $T \in \mathcal{D}$, with $S \subset T$, it is the case that $c(S) \neq c(T) \cap S$. Denote the family of c-maximal sets under the choice domain $\mathcal{D}$ by $M_c^{\mathcal{D}}$.

For any c-maximal set there is no possibility of obtaining its associated elements from any other superset of it included in the domain $\mathcal{D}$, and thus, the study of choice behavior needs to incorporate it.[5] Clearly, all other choice sets in $\mathcal{D}$ can be trivially associated with at least one c-maximal set.

2.2 NP-completeness

We now present an informal introduction to the notion of complexity we use in this paper. For an excellent, detailed and formal account see Garey and Johnson (1979).[6]

The theory of NP-completeness is conventionally centered around *decision problems*. These are problems formulated with a yes-or-no answer. Consequently, we define a decision problem that is the binary analog of the problem of finding a minimal RMR of a choice correspondence c on $\mathcal{D}$ as follows.

Rationalization by Complete Preorders in $\mathcal{D}$ (RCP-$\mathcal{D}$): Given a choice correspondence c on $\mathcal{D}$, can we find $k \leq K$ complete preorders that constitute a rationalization by multiple rationales of c?

Appropriately setting $\mathcal{D}$ or $\mathcal{U}$, and complete preorders or linear orders, we may define the binary problems Rationalization by Complete Preorders in $\mathcal{U}$, RCP-$\mathcal{U}$, Rationalization by Linear Orders in $\mathcal{D}$, RLO-$\mathcal{D}$, and Rationalization by Linear Orders in $\mathcal{U}$, RLO-$\mathcal{U}$.

In this paper we use the proof-by-reduction technique to prove that a particular (decision) problem is NP-complete. That is, to prove that a given problem Π in NP is in fact NP-complete, we show that it contains a known NP-complete problem Π' as a special case.

3 Complexity of rationalization

We start with the classic, structured case, where $\mathcal{D} = \mathcal{U}$ and WARP holds. Let us call this case the *rational procedure*. Then we will study the polar case of the rational procedure.

[5] In the simpler case of single-valued choice functions, the condition to identify the class of c-maximal sets is simply $c(S) \neq c(T)$.

[6] See also Cormen, Leiserson, Rivest, and Stein (2001). Ballester (2004) and Aragones, Gilboa, Postlewaite and Schmeidler (2005) provide introductions in the context of economics.

This is the case where no restriction whatsoever is imposed either on choice behavior, or on the choice domain. At the same time, we will deal with a number of intermediate cases.

3.1 Structured choice behavior: The rational procedure

We start by considering the case where a choice correspondence defined over the universal choice domain $\mathcal{U}$ satisfies WARP. We have already mentioned that this represents a very structured case, as it is well-known that there exists a unique complete preorder relation $\succeq$ rationalizing c.

Now the question arises of how difficult it is to find the rationalizing binary relation $\succeq$. Intuitively, it seems that the high degree of structure of the rational procedure implies that finding the rationale is not a difficult task. This is in fact the case. It is easy to show that the set $|M_c^{\mathcal{U}}|$ is small, as it contains at most n elements. The simplicity of the family of maximal sets allows us to consider very simple algorithms to obtain the rationalization of choice behavior in polynomial time. Consider for instance the following trivial one.

Take $X_0 = X$ and iteratively define $X_k = X_{k-1} \setminus c(X_{k-1})$ if $X_{k-1} \setminus c(X_{k-1}) \neq \emptyset$ holds. Then, the set of maximal elements is the collection of sets $\{X_k\}$, with cardinality equal to or less than n. Clearly, every element is chosen from exactly one set X_k. Then, the rationale can be defined by stating that $x \succeq y$ if and only if $x \in X_l$, $y \in X_j$, with X_l, X_j in the family of c-maximal sets $\{X_k\}$ and $k \leq j$.

We summarize the above in the following observation.

Observation 3.1 *Let the choice correspondence c be a rational procedure and $\mathcal{D} = \mathcal{U}$, then $|M_c^{\mathcal{U}}| \leq n$ and the problem of finding the complete preorder $\succeq$ that rationalizes c is polynomial.*

There are two important remarks to Observation 3.1. First, as a corollary to the above, finding the linear order $\succ$ rationalizing a single-valued choice function defined on $\mathcal{U}$ that satisfies IIA is also polynomial. This follows from the fact that the single-valued case is but a special case of the multi-valued choice correspondence.

Second, the results also hold for arbitrary choice domains $\mathcal{D}$. To find the binary relation rationalizing c when $\mathcal{D} \neq \mathcal{U}$, simply write $x \succeq y$ if and only if $x \in c(A)$ and $y \in A$, $A \in \mathcal{D}$. It is easy to see that WARP guarantees that such a binary relation is a complete preorder.

3.2 Unstructured behavior and unrestricted domain

We now turn to the polar case of the rational procedure where we neither impose structure on choice behavior, nor on the domain of choice sets. This means that, in general, there is not a single binary relation rationalizing choice behavior, but a set of them in the sense of KRS. It is clear that the problem of finding the rationales in this case is a much more demanding task than the one we faced for the rational procedure in the previous section. In fact we show below that the task is demanding to the point of being intractable. That is, we show that finding a minimal RMR in this setting belongs to the class of NP-complete problems, and hence unless P=NP, there is no hope of finding an efficient algorithm that for every choice correspondence c gives a minimal RMR in a reasonable time frame.

We can now state our first NP-complete result.

Theorem 3.2 *RCP-$\mathcal{D}$ is NP-complete.*

It can be shown that the c used in the proof is single-valued and the binary relations defined from the partition of the graph are linear orders. Hence, the following corollary to (the proof of) Theorem 3.2 is immediate.

Corollary 3.3 *RLO-$\mathcal{D}$ is NP-complete.*

Theorem 3.2 (and Corollary 3.3) show that the conjunction of (i) unstructured choice behavior and (ii) unrestricted choice domain lead to the NP-completeness of the problem of rationalization. But are the two conditions required to get the intractability result? In principle, it could be that the real difficulty in finding a minimal RMR is completely triggered by choice behavior per se, or it could be that it is the interplay of behavior and domain that drives the result. Theorem 3.4 below suggests that in the case of single-valued choice functions it is the interplay of both, unstructured behavior and unrestricted domain, that triggers the intractability of finding the rationales.

Theorem 3.4 *RLO-$\mathcal{U}$ is quasi-polynomially bounded.*

Theorem 3.4 suggests that RLO-$\mathcal{U}$ is not NP-complete.[7] Otherwise, Theorem 3.4 would imply that *all* NP-complete problems are quasi-polynomially bounded. However, in spite of continuous efforts such a bound has never been found for NP-complete problems. Indeed, there is the strong conviction this will never happen.

The question with regard to choice correspondences defined on $\mathcal{U}$, however, remains open. The naive algorithm used in Theorem 3.4 is not quasi-polynomially bounded in this case. It is not difficult to see that the number of possible complete preorders is upper bounded by $n!^2$. The problem arises because, in the case of complete preorders, the $n-1$ bound on the maximum number of preference relations to check for rationalization does not hold. In fact, it is straightforward to see that the bound turns out to be 2^n. This implies that the naive algorithm for choice correspondences cannot be quasi-polynomially bounded, having an exponential order of magnitude. Therefore, it may well be the case that with choice correspondences, the practical difficulty in finding a minimal RMR is triggered by choice behavior per se.

The challenge for future research is then to understand the driving force of the complexity of rationalization, and use this understanding to find specific algorithms that behave well under certain circumstances. In the next section we start this task by drawing a connection with graph theory. This is especially useful since there is a wealth of algorithms for graph problems that may solve the problem of rationalization of certain choice structures.

4 On the structure of the complexity of rationalization

4.1 Rationalization and graph theory

The following binary relations capture two fundamental properties on the structural relation of maximal sets.

Definition 4.1 *Let $A, B \in M_c^{\mathcal{D}}$, $A \to B$ if and only if $c(A) \cap B \notin \{\emptyset, c(B) \cap A\}$.*

This first binary relation $\to$ states the conditions under which a set A *blocks* another set B, in the sense that A and B cannot be rationalized by a binary relation $\succeq$ writing $c(A) \succeq c(B)$. Define $H_A = A \setminus c(A)$, $A \in \mathcal{D}$. Then, when c is single-valued, $\to$ reduces to: $A \to B$ if and only if $c(A) \in H_B$.

Definition 4.2 *Let $A, B \in M_c^{\mathcal{D}}$, $A - B$ if and only if $c(A) \cap c(B) \neq \emptyset$.*

[7] The naive algorithm of Theorem 3.4 does not necessarily behave well in unrestricted choice domains $\mathcal{D}$. The reason being that in this case the input size need not be as high as 2^n, but for example, it could be n. This gives an exponential order of magnitude for the naive algorithm.

$A - B$ means that sets A and B are linked in the rationalization problem. That is, whenever $A - B$, if A and B are to be rationalized by the same binary relation $\succeq$, then it must be that $c(A) \sim c(B)$.

Finally, we use the above two binary relations to define an oriented cycle.

Oriented Cycle: The collection $\{A_t\}_{t=1}^n \in M_c^{\mathcal{D}}$, $n \geq 2$, is an oriented cycle if

1. $A_1 = A_n$,
2. for every $i \in \{1, \ldots, n-1\}$, either $A_i \to A_{i+1}$, or $A_i - A_{i+1}$, and
3. there is $j \in \{1, \ldots, n-1\}$ such that $A_j \to A_{j+1}$.

We are now in a position to introduce a graph theory problem over the space of c-maximal sets.

Minimal Non-Oriented Partition (NOP): A partition of $M_c^{\mathcal{D}}$ $\{V_p\}_{p=1,\ldots,P}$ is said to be a non-oriented partition if for every class V_p there is no oriented cycle. It is said to be minimal if any other NOP has at least P classes.

A (minimal) NOP is constructed over the set $M_c^{\mathcal{D}}$ according to the binary relations $\to$ and $-$. Hence, a class V_p in the partition has a clear interpretation: all the sets in the class V_p can be rationalized through a complete preorder $\succeq_p$. Then, an NOP gives information on which choice problems can be rationalized together.

The following theorem establishes that finding a minimal RMR is equivalent to finding a minimal NOP. This opens the possibility of drawing upon the established algorithm knowledge on graph theory problems.

Theorem 4.3 *Let c be a choice correspondence:*
-If $\{\succeq_p\}_{p=1,\ldots,P}$ is a minimal RMR, then there is a minimal NOP $\{V_p\}_{p=1,\ldots,P}$.
-If $\{V_p\}_{p=1,\ldots,P}$ is a minimal NOP, then there is a minimal RMR $\{\succeq_p\}_{p=1,\ldots,P}$.

The significance of Theorem 4.3 is best appreciated in the case of single-valued choice functions. Recall that when c is single-valued, $A \to B$ if and only if $c(A) \in H_B$, $A, B \in M_c^{\mathcal{D}}$. That is, A blocks B if and only if the chosen element in A belongs to B and, at the same, time this element is not chosen in B. It is clear that such a case is inconsistent with a linear order rationalizing B and writing $c(A) \succ c(B)$. Also, note that $A - B$ if and only if $c(A) = c(B)$. Then the relation between minimal RMRs and minimal NOPs simplifies considerably. This is because for any three sets $A_1, A_2, B \in M_c^{\mathcal{D}}$, whenever $c(A_1) = c(A_2)$, then $A_1 \to B$ if and only if $A_2 \to B$. This implies that, whenever there is an oriented cycle, there is a cycle composed of elements related only through $\to$. Hence the analysis can obviate the equivalence relation $-$, and focus on $(M_c^{\mathcal{D}}, \to)$. The latter is simply an standard directed graph, and the structure defined as an oriented cycle reduces to the standard notion of a directed cycle. There is an immense literature on graphs without directed cycles, typically known as directed acyclic graphs (DAGs).[8] This is especially important since Theorem 4.3 guarantees that there is much to gain from the results in this literature. Hence, our problem reduces to find a minimal partition into DAGs.

[8]See, e.g., Cormen, Leiserson, Rivest, and Stein (2001).

4.2 An example

The study of concrete choice procedures appears particularly appealing. It is likely that the structure inherent to specific choice procedures allows for the tractability of rationalization. Here we provide an example that draws on the previous subsection.

Manzini and Mariotti (2007) study the nature of choice functions defined on $\mathcal{U}$ that can be rationalized by sequentially applying a set of asymmetric binary relations in a fixed order. Among other results, they provide a full characterization of the case when a choice function is sequentially rationalized by two asymmetric binary relations. Such a choice function is called a Rational Shortlist Method (RSM). Their characterization makes use of the classical property of expansion.

Expansion: If $x \in c(A)$ and $x \in c(B)$, $A, B \in \mathcal{U}$, then $x \in c(A \cup B)$.

It is not difficult to observe that the set $M_c^{\mathcal{U}}$ shrinks considerably whenever this property holds. Clearly, for each element x in X there is at most one element in $M_c^{\mathcal{U}}$ for which x is the chosen element, namely, the union of all subsets A for which x is chosen. Denote the latter by $M(x)$. Then the problem of finding a minimal RMR here reduces to finding a minimal partition into DAGs over the universal set of alternatives X according to $\rightleftharpoons$, where for every $x \neq y$,

$$x \rightleftharpoons y \text{ if and only if } M(x) \to M(y) \text{ if and only if } x \in M(y)$$

Hence the rationalization of RSMs through multiple rationales turns out to be a computationally tractable problem.

5 Final remarks

We have used the tools of theoretical computer science to study the complexity of finding the preference relations that rationalize choice behavior. The question of rationalizability of choice behavior has played a central role in economics. However, surprisingly enough, virtually no attention has been given to the practical problem of computing the rationales.

We have shown that, in the classical case, when the weak axiom of revealed preference holds, finding the preference relation rationalizing choice behavior is easy. There are polynomial time algorithms that compute the rationale quickly. On the other hand, when we neither impose any restriction on choice behavior nor on the domain, we have shown that the problem of rationalization is NP-complete. Therefore, there is little hope of finding an efficient algorithm bounded above by a a polynomial function. Furthermore, we have shown that in the case of single-valued choice functions, it is the conjunction of unstructured choice and unrestricted domain that drives the intractability result. Under the universal domain, the problem of finding a minimal book is quasi-polynomially bounded. On the other hand, we argue that in the choice correspondences case, it may well be the case that the difficulty in finding a minimal book is triggered by choice behavior per se.

We then turned to trying to better understand the complexity of rationalization. To this end we identified two binary relations over choice sets that capture part of the essence of rationalization. Furthermore, these binary relations define a problem in graph theory that is equivalent to the problem of rationalization. This is particularly interesting since the complexity issues have attracted a great deal of attention in graph theory. The equivalence result provided allows for the searching of existing algorithms in graph theory that can be used for the problem of rationalization.

Apart from the literature on rationalization, our results relate to two other strands in the literature. First, the problem of rationalization can be read as the problem of transmission

of information (choice behavior in different situations), given a specific grammar (complete preorders a la KRS). Under this interpretation, our paper is related to the economics literature on language (see Rubinstein, 2000). Rubinstein stresses the importance of binary relations to natural language. In this sense, our results establish that there are practical limitations to the design of a grammar with the ability to transmit any kind of information. Several questions arise. Does the structure of natural speech imply the existence of a collection of complete preorders computable in polynomial time? What, if anything, is lost in the transmission of information, if the problem of constructing a grammar is upper bounded by a polynomial time algorithm?

Second, an immediate conclusion from our results is that the more structured behavior is, the easier it is to rationalize it in practice. That is, rationality makes things easier. This type of observation has been recognized in the bounded rationality literature from different perspectives. Tversky and Simonson (1993) note that the standard maximization problem is hard to beat in terms of its simplicity of formulation. It is most likely that any descriptive bounded rationality model is condemned to involve a more cumbersome formulation. Also, Salant (2003) shows that the rational procedure requires the least memory possible and that the automatation required to compute the rational procedure is the smallest possible.

References

[1] Aragones, E., I. Gilboa, A. Postlewaite, and D. Schmeidler (2005), "Fact-Free Learning," *American Economic Review*, 95:1355-1368

[2] Arrow, K.J. (1959), "Rational Choice Functions and Orderings," *Economica*, 26:121-127.

[3] Ballester, C. (2004), "NP-completeness in Hedonic Games," *Games and Economic Behavior*, 49:1-30.

[4] Cormen, T.H., C.E. Leiserson, R.L. Rivest, and C. Stein (2001), *Introduction to Algorithms.* MIT Press.

[5] Garey, M.R. and D.S. Johnson (1979), *Computers and Intractability: A Guide to the Theory of NP-Completeness.* Freeman.

[6] Kalai, G., A. Rubinstein, and R. Spiegler (2002), "Rationalizing Choice Functions by Multiple Rationales," *Econometrica*, 70:2481-88.

[7] Manzini, P. and M. Mariotti (2007), "Sequentially Rationalizable Choice," *American Economic Review*, 97:1824-1839.

[8] Moulin, H. (1985), "Choice Functions over a Finite Set: A Summary," *Social Choice and Welfare*, 2:147-160.

[9] Ok, E.A. (2005), "Independence of Irrelevant Alternatives and Individual Choice," mimeo, New York University.

[10] Richter, M.K. (1966), "Revealed Preference Theory," *Econometrica*, 34:635-645.

[11] Rubinstein, A. (2000), *Economics and Language*, Cambridge University Press.

[12] Salant, Y. (2003), "Limited Computational Resources Favor Rationality," mimeo, Hebrew University.

[13] Sen, A.K. (1971), "Choice Functions and Revealed Preferences," *Review of Economic Studies*, 38:307-317.

[14] Suzumura, K. (1983), *Rational Choice, Collective Decisions, and Social Welfare*, Cambridge University Press, Cambridge.

[15] Tversky, A. and I. Simonson (1993), "Context-dependent Preferences," *Management Science*, 39:1179-1189.

[16] Uzawa, H. (1957), "Notes on Preference and the Axiom of Choice," *Annals of the Institute of Statistical Mathematics*, 8:35-40.

[17] Xu, Y. and L. Zhou (2007) "Rationalizability of Choice Functions by Game Trees," *Journal of Economic Theory*, 134:548–556.

Jose Apesteguia
Department of Economics
Universitat Pompeu Fabra
Email: jose.apesteguia@upf.edu

Miguel A. Ballester
Department of Economics
Universitat Autònoma de Barcelona
Email: MiguelAngel.Ballester@uab.es

Alternatives to Truthfulness are Hard to Recognize

Vincenzo Auletta, Paolo Penna, Giuseppe Persiano and Carmine Ventre

Abstract

The central question in mechanism design is how to implement a given social choice function. One of the most studied concepts is that of *truthful* implementations in which truth-telling is always the best response of the players. The Revelation Principle says that one can focus on truthful implementations without loss of generality (if there is no truthful implementation then there is no implementation at all). Green and Laffont [1] showed that, in the scenario in which players' responses can be *partially verified*, the revelation principle holds only in some particular cases.

When the Revelation Principle does not hold, non-truthful implementations become interesting since they might be the only way to implement a social choice function of interest. In this work we show that, although non-truthful implementations may exist, they are hard to find. Namely, it is NP-hard to decide if a given social choice function can be implemented in a non-truthful manner, or even if it can be implemented at all. This is in contrast to the fact that truthful implementability can be recognized efficiently, even when partial verification of the agents is allowed. Our results also show that there is no "simple" characterization of those social choice functions for which it is worth looking for non-truthful implementations.

1 Introduction

Social choice theory deals with the fact that individuals (agents) have different preferences over the set of possible alternatives or outcomes. A social choice function maps these preferences into a particular outcome, which is not necessarily the one preferred by the agents. The main difficulty in implementing a social choice function stems from the fact that agents can *misreport* their preferences. Intuitively speaking, a social choice function can be implemented if there is a method for selecting the desired outcome which cannot be manipulated by *rational* agents. By 'desired outcome' we mean the one specified by the social choice function applied to the *true* agents' preferences.

More precisely, each agent has a *type* which specifies the utility he derives if some outcome is selected. When agents are also endowed with payments, we consider agents with quasi linear utility: the type specifies the gross utility and the agent's utility is the sum of gross utility and payment received. In either case, a rational agent reports a type so to maximize his own utility and the reported type must belong to a *domain* consisting of all possible types. In the case of *partially verifiable* information, the true type of an agent further restricts the set of types that he can possibly report [1].

One of the most studied solution concepts is that of *truthful* implementations in which agents always maximize their utilities by truthfully reporting their types. The *Revelation Principle* says that one can focus on truthful implementations without loss of generality: A social choice function is implementable if and only if it has a truthful implementation. Green and Laffont [1] showed that, in the case of partially verifiable information, the Revelation Principle holds only in some particular cases. When the Revelation Principle does not hold, non-truthful implementations become interesting since they might be the only way to implement a social choice function of interest. Although a non-truthful implementation may induce some agent to misreport his type, given that he reports the type maximizing his utility, it is still possible to compute the desired outcome "indirectly". Singh and Wittman [3] observed that the Revelation Principle fails in several interesting cases and show sufficient conditions for the existence of non-truthful implementations.

1.1 Our contribution

In this work, we study the case in which the declaration of an agent can be partially verified. We adopt the model of Green and Laffont [1] in which the ability to partially verify the declaration of an agent is encoded by a *correspondence* function M: $M(t)$ is the set of the possible declarations of an agent of type t. Green and Laffont [1] characterized the correspondences for which the Revelation Principle holds; that is, correspondences M for which a social choice function is either truthfully implementable or not implementable at all.

We show that although non-truthful implementations may exist, they are hard to find. Namely, it is NP-hard to decide if a given social choice function can be implemented for a given correspondence in a non-truthful manner. This is in contrast to the fact that it is possible to efficiently decide whether a social choice function can be truthfully implemented for a given correspondence. Our results show that there is no "simple" characterization of those social choice functions that violate the Revelation Principle. These are the social choice functions for which it is worth looking for non-truthful implementations since this might be the only way to implement them.

We prove these negative results for a very restricted scenario in which we have only one agent and at most two possible outcomes, and the given function does not have truthful implementations. We give hardness proofs both for the case in which payments are not allowed and the case in which payments are allowed and the agent has quasi linear utility.

In general payments are intended as a tool for enlarging the class of social choice functions that can be implemented. We find that there is a rich class of correspondences for which it is NP-hard to decide if a social choice function can be implemented *without* payments, while for the same correspondences it is trivial to test truthful implementability with payments via the approach in [3]. Finally, we complement our negative results by showing a class of correspondences for which there is an efficient algorithm for deciding whether a social choice function can be implemented.

We note that the characterization of Green and Laffont [1] has no direct implication in our results. Indeed, the property characterizing the Revelation Principle can be tested efficiently. Moreover, when the Revelation Principle does not hold, we only know that there exists *some* social choice function which is only implemented in a non-truthful manner. Hence, we do not know if the social choice function of interest can be implemented or not. Note that this question can be answered efficiently when the Revelation Principle holds since testing the existence of truthful implementations is computationally easy.

Road map. We introduce the model with partial verification by Green and Laffont [1] in Section 2. The case with no payments is studied in Section 3. Section 4 presents our results for the case in which payments are allowed and the agent has quasi linear utility. We draw some conclusions in Section 5.

2 The Model

The model considered in this work is the one studied by Green and Laffont [1] who considered the so called principal-agent scenario. Here there are two players: the agent, who has a type t belonging to a domain D, and the principal who wants to compute a social choice function $f : D \to \mathcal{O}$, where $\mathcal{O}$ is the set of possible outcomes. The quantity $t(X)$ denotes the *utility* that an agent of type t assigns to outcome $X \in \mathcal{O}$.

The agent observes his type $t \in D$ and then transmits some message $t' \in D$ to the principal. The principal applies the outcome function $g : D \to \mathcal{O}$ to t' and obtains outcome $X = g(t')$. We stress that the principal fixes the outcome function g in advance and then the agent *rationally* reports t' so to maximize his utility $t(g(t'))$. Even though the principal does not exactly know the type of the agent, it is reasonable to assume that some *partial* information on the type of the agent is available. Thus the agent is restricted to report a type t' in a set $M(t) \subseteq D$, which is specified by a *correspondence* function $M : D \to 2^D$. We will only consider correspondences $M(\cdot)$ for which truth-telling is always an option; that is, for all $t \in D$, $t \in M(t)$. Notice that the case in which the principal has no information (no verification is possible) corresponds to setting $M(t) = D$ for all t.

Definition 1 ([1]) *A mechanism (M, g) consists of a correspondence $M : D \to 2^D$ and an* outcome function *$g : D \to \mathcal{O}$. The outcome function g induces a best response rule $\phi_g : D \to D$ defined by $\phi_g(t) \in \arg\max_{t' \in M(t)}\{t(g(t'))\}$. If $t \in \arg\max_{t' \in M(t)}\{t(g(t'))\}$ then we set $\phi_g(t) = t$.*

The correspondence M can be represented by a directed graph $\mathcal{G}_M$ (which we call the *correspondence graph*) defined as follows. Nodes of $\mathcal{G}_M$ are types in the domain D and an edge (t, t'), for $t \neq t'$, exists if and only if $t' \in M(t)$. We stress that the correspondence graph of M does not contain self-loops, even though we only consider correspondences M such that $t \in M(t)$ for all $t \in D$. We will often identify the correspondence M with its correspondence graph $\mathcal{G}_M$ and say, for example, that a correspondence is acyclic meaning that its correspondence graph is acyclic. Sometimes it is useful to consider a weighted version of graph $\mathcal{G}_M$. Specifically, for a function $g : D \to \mathcal{O}$, we define $\mathcal{G}_{M,g}$ to be the weighted version of graph $\mathcal{G}_M$ where edge (t, t') has weight $t(g(t)) - t(g(t'))$.

We study the class of M-implementable social choice functions $f : D \to \mathcal{O}$.

Definition 2 ([1]) *An outcome function $g : D \to \mathcal{O}$* M-implements *social choice function $f : D \to \mathcal{O}$ if for all $t \in D$ $g(\phi_g(t)) = f(t)$ where $\phi_g(\cdot)$ is the best response rule induced by g. A social choice function $f : D \to \mathcal{O}$ is* M-implementable *if and only if there exists an outcome function $g : D \to \mathcal{O}$ that M-implements f.*

The social choice functions that can be truthfully M-implemented are of particular interest.

Definition 3 ([1]) *An outcome function $g : D \to \mathcal{O}$* truthfully M-implements *social choice function $f : D \to \mathcal{O}$ if g M-implements f and $\phi_g(t) = t$ for all $t \in D$. A social choice function $f : D \to \mathcal{O}$ is* truthfully M-implementable *if and only if there exists an outcome function $g : D \to \mathcal{O}$ that truthfully M-implements f.*

The classical notions of *implementation* and of *truthful implementation* are obtained by setting $M(t) = D$ for all $t \in D$. Actually in this case the two notions of implementable social choice function and of truthfully implementable social choice function coincide due to the well-known revelation principle.

Theorem 4 (The Revelation Principle) *If no verification is possible (that is, $M(t) = D$ for all $t \in D$), a social choice function is implementable if and only if it is truthfully implementable.*

The Revelation Principle does not necessarily hold for the notion of M-implementation and of truthful M-implementation. Green and Laffont [1] indeed give a necessary and sufficient condition on M for the revelation principle to hold. More precisely, a correspondence M satisfies the *Nested Range Condition* if the following holds: for any $t_1, t_2, t_3 \in D$ if $t_2 \in M(t_1)$ and $t_3 \in M(t_2)$ then $t_3 \in M(t_1)$.

Theorem 5 (Green-Laffont [1]) *If M satisfies the NRC condition then a social choice function f is M-implementable if and only if f is M-truthfully implementable. If M does not satisfy the NRC condition then there exists an M-implementable social choice function f that is not truthfully M-implementable.*

Besides its conceptual beauty, the Revelation Principle can also be used in some cases to decide whether a given social choice function f is M-implementable for a given correspondence M. Indeed, if the Revelation Principle holds for correspondence M, the problem of deciding M-implementability is equivalent to the problem of deciding truthful M-implementability which, in turn, can be efficiently decided.

Theorem 6 *There exists an algorithm running in time polynomial in the size of the domain that, given a social choice function f and a correspondence M, decides whether f is truthfully M-implementable.*

PROOF. To test truthful M-implementability of f we consider graph $\mathcal{G}_{M,f}$ where edge (t, t') has weight $t(f(t)) - t(f(t'))$. Then it is obvious that f is M-truthful implementable if and only if no edge of $\mathcal{G}_{M,f}$ has negative weight. □

3 Hardness of the Implementability problem

In this section we prove that the following problem is NP-hard.

Problem 1 *The* IMPLEMENTABILITY *problem is defined as follows.*

INPUT: *domain D, outcome set $\mathcal{O}$, social choice function $f : D \to \mathcal{O}$ and correspondence M.*

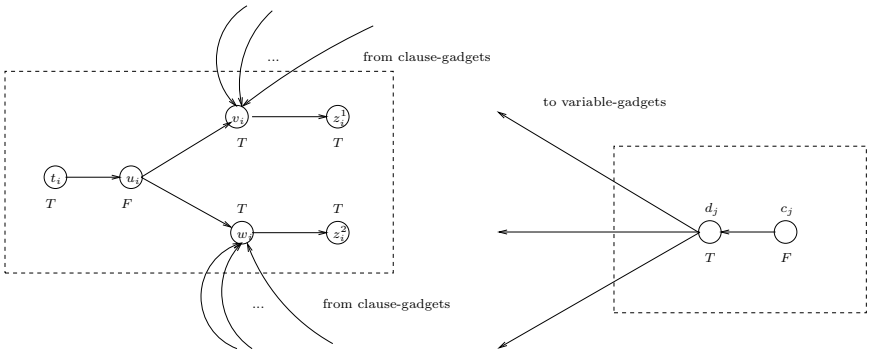
TASK: *decide whether there exists an outcome function g that M-implements f.*

The following lemma, whose proof is immediate, gives sufficient conditions for an outcome function g to M-implement social choice function f.

Lemma 7 *For outcomes $\mathcal{O} = \{T, F\}$, if the following conditions are satisfied for all $a \in D$ then outcome function g M-implements social choice function f.*

1. *If $f(a) = T$ and $a(T) < a(F)$ then, for all $v \in M(a)$, we have $g(v) = T$.*
2. *If $f(a) = F$ and $a(T) < a(F)$ then, there exists $v \in M(a)$ such that $g(v) = F$.*
3. *If $f(a) = T$ and $a(T) > a(F)$ then, there exists $v \in M(a)$ such that $g(v) = T$.*
4. *If $f(a) = F$ and $a(T) > a(F)$ then, for all $v \in M(a)$, we have $g(v) = F$.*

The reduction. We reduce from 3SAT. Let Φ a Boolean formula in 3-CNF over the variables $x_1, \cdots, x_n$ and let $C_1, \cdots, C_m$ be the clauses of Φ. We construct D, $\mathcal{O}$, M and $f : D \to \mathcal{O}$ such that f is M-implementable if and only if Φ is satisfiable. We set $\mathcal{O} = \{T, F\}$. We next construct a correspondence graph $\mathcal{G}_M$ representing M. We will use variable gadgets (one per variable) and clause gadgets (one per clause).



(a) The variable gadget (b) The clause gadget

Figure 1: Gadgets used in the reduction.

The variable gadget for the variable x_i is depicted in Figure 1(a). Each variable x_i of the formula Φ adds six new types to the domain D of the agent, namely, t_i, u_i, v_i, w_i, z_i^1 and z_i^2 satisfying the following relations:

$$t_i(F) > t_i(T), \quad (1)$$
$$u_i(F) > u_i(T), \quad (2)$$
$$v_i(T) > v_i(F), \quad (3)$$
$$w_i(T) > w_i(F). \quad (4)$$

The labeling of the vertices defines the social choice function f; that is, $f(t_i) = T, f(v_i) = T, f(w_i) = T, f(z_i^1) = T, f(z_i^2) = T,$ and $f(u_i) = F$. Directed edges of the gadget describe the correspondence M (rather the correspondence graph). Thus, for example, $M(t_i) =$

$\{t_i, u_i\}$ and $M(u_i) = \{u_i, v_i, w_i\}$. Nodes v_i and w_i have incoming edges from the clause gadgets. The role of these edges will be clear in the following.

We observe that (1) implies that the social choice function f is not truthfully M-implementable. Indeed t_i prefers outcome $F = f(u_i)$ to $T = f(t_i)$ and $u_i \in M(t_i)$. Moreover, by Lemma 7, for any outcome function g implementing f we must have $g(t_i) = g(u_i) = T$. On the other hand, since $f(u_i) = F$ it must be the case that any g that M-implements f assigns outcome F to *at least* one node in $M(u_i) \setminus \{u_i\}$. Intuitively, the fact that every outcome function g that M-implements f must assign F to at least one between v_i and w_i corresponds to assigning "false" to respectively literal x_i and $\bar{x}_i$.

The clause gadget for clause C_j of Φ is depicted in Figure 1(b). Each clause C_j adds types c_j and d_j to the domain D of the agent such that

$$c_j(T) > c_j(F), \tag{5}$$
$$d_j(T) > d_j(F). \tag{6}$$

As before the labeling defines the social choice function f and we have $f(d_j) = T$ and $f(c_j) = F$. Moreover, directed edges encode correspondence M. Besides the directed edge (c_j, d_j), the correspondence graph contains three edges directed from d_j towards the three variable gadgets corresponding to the variables appearing in the clause C_j. Specifically, if C_j contains the literal x_i then d_j has an outgoing edge to node v_i. If C_j contains the literal $\bar{x}_i$ then d_j has an outgoing edge to node w_i. Similarly to the variable gadget, we observe that (5) implies that for any g M-implementing f it must be $g(d_j) = F$. Therefore, for g to M-implement f it must be the case that, for at least one of the neighbors a of d_j from a variable gadget, we have $g(a) = T$. We will see that this happens if and only if the formula Φ is satisfiable. This concludes the description of the reduction.

We next prove that the reduction is correct. Suppose that Φ is satisfiable, let τ be a satisfying truth assignment and let g be the outcome function defined as follows. For the i-th variable gadget we set $g(t_i) = g(u_i) = g(z_i^1) = g(z_i^2) = T$. Moreover, if x_i is true in τ, then we set $g(v_i) = T$ and $g(w_i) = F$; otherwise se set $g(v_i) = F$ and $g(w_i) = T$. For the j-th clause gadget, we set $g(d_j) = g(c_j) = F$.

Thus, to prove that the outcome function produced by our reduction M-implements f, it is sufficient to show for each type a the corresponding condition of Lemma 7 holds. We prove that conditions hold only for $a = u_i$ and $a = d_j$, the other cases being immediate. For u_i we have to verify that Condition 2 of Lemma 7 holds. Since τ is a truth assignment, for each i vertex u_i has a neighbor vertex for which the outcome function g gives F. For d_j we have to verify that Condition 3 of Lemma 7 holds. Since τ is a satisfying truth assignment, for each j there exists at least one literal of C_j that is true in τ; therefore, vertex d_j has a neighbor vertex for which the outcome function g gives T.

Conversely, consider an outcome function g which M-implements the social choice function f. This means that, for each clause C_j, d_j is connected to at least one node, call it a_j, from a variable gadget such that $g(a_j) = T$. Then the truth assignment that sets to true the literals corresponding to nodes $a_1, \cdots, a_m$ (and gives arbitrary truth value to the other variables) satisfies the formula.

The following theorem follows from the above discussion and from the observation that the reduction can be carried out in polynomial time and the graph we constructed is acyclic with maximum outdegree 3.

Theorem 8 *The* Implementability *Problem is NP-hard even for outcome sets of size* 2 *and acyclic correspondences of maximum outdegree* 3*.*

3.1 Corrrespondences with outdegree 1

In this section, we study correspondences of outdegree 1.

We start by reducing the problem of finding g that M-implements f, for the case in which $\mathcal{G}_M$ is a line, to the problem of finding a satisfying assignment for a formula in 2CNF (that is every clause has at most 2 literals). We assume $D = \{t_1, \cdots, t_n\}$, $\mathcal{O} = \{o_1, \cdots, o_m\}$ and that, for $i = 2, \cdots, n$, $M(t_i) = \{t_i, t_{i-1}\}$ and $M(t_1) = \{t_1\}$. We construct a formula Φ in 2CNF in the following way. The formula Φ has the variables x_{ij} for $1 \leq i \leq n$ and $1 \leq j \leq m$. The intended meaning of variable x_{ij} being set to true is that $g(t_i) = o_j$. We will construct Φ so that every truth assignment that satisfies Φ describes g that M-implements f. We do so by considering the following clauses:

1. Φ contains clauses $(x_{if(t_i)} \vee x_{i-1f(t_i)})$, for $i = 2, \cdots, n$, and clause $x_{1f(t_1)}$.

   These clauses encode the fact that for g to M-implement f it must be the case that there exists at least one neighbor a of t_i in $\mathcal{G}_M$ such that $g(a) = f(t_i)$.

2. Φ contains clauses $(x_{ij} \rightarrow \overline{x}_{ik})$, for $i = 1, \cdots, n$ and for $1 \leq k \neq j \leq m$.

   These clauses encode the fact that g assigns at most one outcome to t_i.

3. Φ contains clauses $(x_{if(t_i)} \rightarrow \overline{x}_{i-1k})$ for all $i = 2, \cdots, n$ and for all k such that $t_i(o_k) > t_i(f(t_i))$.

   These clauses encode the fact that if g M-implements f and $g(t_i) = f(t_i)$ then agent of type t_i does not prefer $g(t_{i-1})$ to $g(t_i)$. Therefore, in this case t_i's best response is t_i itself.

4. Φ contains clauses $(x_{i-1f(t_i)} \rightarrow \overline{x}_{ik})$ for all $i = 2, \cdots, n$ and for all k such that $t_i(o_k) \geq t_i(f(t_i))$.

   These clauses encode the fact that if g M-implements f and $g(t_{i-1}) = f(t_i)$ then agent of type t_i does not prefer $g(t_i)$ to $g(t_{i-1})$. Therefore, in this case t_i's best response is t_{i-1}.

It is easy to see that Φ is satisfiable if and only if f is M-implementable. The above reasoning can be immediately extended to the case in which each node of $\mathcal{G}_M$ has outdegree at most 1 (that is $\mathcal{G}_M$ is a collection of cycles and paths). We thus have the following theorem.

Theorem 9 *The* Implementability *Problem can be solved in time polynomial in the sizes of the domain and of the outcome sets for correspondences of maximum outdegree* 1.

4 Implementability with quasi linear utility

In this section we consider mechanisms with payments; that is, the mechanism picks an outcome and a payment to be transferred to the agent, based on the reported type of the agent. Therefore a mechanism is now a pair (g, p) where g is the outcome function and $p : D \rightarrow \mathbb{R}$ is the payment function. We assume that the agent has quasi linear utility.

Definition 10 *A mechanism (M, g, p) for an agent with quasi-linear utility is a triplet where $M : D \rightarrow 2^D$ is a correspondence, $g : D \rightarrow D$ is an outcome function, and $p : D \rightarrow \mathbb{R}$ is a payment function.*

The mechanism defines a best-response function $\phi_{(g,p)} : D \rightarrow D$ where $\phi_{(g,p)}(t) \in \arg\max_{t' \in M(t)}\{t(g(t')) + p(t')\}$. If $t \in \arg\max_{t' \in M(t)}\{t(g(t')) + p(t')\}$ then we set $\phi_g(t) = t$.

Definition 11 *The pair (g,p) M-implements social choice function $f : D \to \mathcal{O}$ for an agent with quasi-linear utility if for all $t \in D$, $g(\phi_{(g,p)}(t)) = f(t)$.*

The pair (g,p) truthfully M-implements social choice function f for an agent with quasi-linear utility if (g,p) M-implements f and, for all $t \in D$, $\phi_{(g,p)}(t) = t$.

In the rest of this section we will just say that (g,p) M-implements (or truthfully M-implements) f and mean that M-implementation is for agent with quasi-linear utility.

Testing truthful M-implementability of a social choice function f can be done in time polynomial in the size of the domain by using the following theorem that gives necessary and sufficient conditions. The proof is straightforward from the proof of [2] (see also [4]).

Theorem 12 *Social choice function f is truthfully M-implementable if and only if $\mathcal{G}_{M,f}$ has no negative weight cycle.*

As in the previous case when payments were not allowed, if M has the NRC property then the Revelation Principle holds and the class of M-implementable social choice functions coincides with the class of truthfully M-implementable social choice functions. We next ask what happens for correspondences M for which the NRC property does not hold. Our answer is negative as we show that the following problem is NP-hard.

Problem 2 *The* Quasi-Linear Implementability *problem is defined as follows.*

Input: *domain D, outcome set $\mathcal{O}$, social choice function $f : D \to \mathcal{O}$ and correspondence M.*

Task: *decide whether there exists (g,p) that M-implements f.*

We start with the following technical lemma.

Lemma 13 *Let M be a correspondence and let f be a social choice function for which correspondence graph has a negative-weight cycle $t \to t' \to t$ of length 2. If (g,p) M-implements f then*

$$\{\phi_{(g,p)}(t), \phi_{(g,p)}(t')\} \not\subseteq \{t, t'\}.$$

Proof. Let us assume for sake of contradiction that (g,p) M-implements f and that

$$\{\phi_{(g,p)}(t), \phi_{(g,p)}(t')\} \subseteq \{t, t'\}. \tag{7}$$

Since cycle $C := t \to t' \to t$ has weight

$$t(f(t)) - t(f(t')) + t'(f(t')) - t'(f(t)) < 0 \tag{8}$$

then $f(t) \neq f(t')$. Therefore, since (g,p) M-implements f, it holds $\phi_{(g,p)}(t) \neq \phi_{(g,p)}(t')$ and thus (7) implies that $\{\phi_{(g,p)}(t), \phi_{(g,p)}(t')\} = \{t, t'\}$.

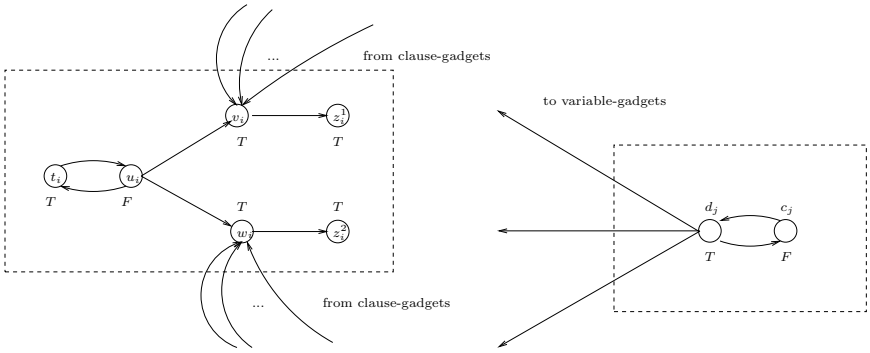
Suppose that $\phi_{(g,p)}(t) = t'$ and thus $\phi_{(g,p)}(t') = t$. Then for (g,p) to M-implement f it must be the case that $g(t) = f(t')$, $g(t') = f(t)$. But then the payment function p must satisfy both the following:

$$\begin{aligned} p(t') + t(f(t)) &\geq p(t) + t(f(t')), \\ p(t) + t'(f(t')) &\geq p(t') + t'(f(t)), \end{aligned}$$

which contradicts (8). The same argument can be used for the case $\phi_{(g,p)}(t) = t$ and $\phi_{(g,p)}(t') = t'$. □

The reduction. We are now ready to show our reduction from 3SAT to the Quasi-Linear Implementability problem. The reduction is similar in spirit to the one of the previous section. We start from a Boolean formula Φ in conjunctive normal form whose clauses contain exactly 3 literals and we construct a domain D, a set of outcomes $\mathcal{O}$, a social choice function f, and a correspondence M such that there exists (g, p) that M-implements f if and only if Φ is satisfiable.

We set $\mathcal{O} = \{T, F\}$ and fix constants $0 < \beta < \delta$. Let $x_1, \ldots, x_n$ be the variables and $C_1, \ldots, C_m$ be the clauses of Φ. The reduction uses two different gadgets: variable gadgets and clause gadgets.



(a) The variable gadget (b) The clause gadget

Figure 2: Gadgets used in the reduction.

We have one variable gadget for each variable; the gadget for x_i is depicted in Figure 2(a) where the depicted edges are edges of $\mathcal{G}_M$. Each variable x_i of the formula Φ adds six new types to the domain D: $t_i, u_i, v_i, w_i, z_i^1$, and z_i^2 satisfying the following two non-contradicting inequalities:

$$t_i(T) - t_i(F) < u_i(T) - u_i(F), \tag{9}$$

$$u_i(T) - u_i(F) = \beta. \tag{10}$$

Nodes v_i and w_i have incoming edges from the clause gadgets. The role of these edges will be clear in the following. The labeling of the nodes describes the social choice function f to be implemented. More precisely, we have that $f(t_i) = f(v_i) = f(w_i) = f(z_i^1) = f(z_i^2) = T$ and $f(u_i) = F$.

We observe that, by (9), cycle $C := t_i \to u_i \to t_i$ has negative weight. Moreover, since $\phi_{(g,p)}(t_i) \in M(t_i) = \{t_i, u_i\}$, by Lemma 13, it must be the case that $\phi_{(g,p)}(u_i) \notin \{t_i, u_i\}$. Therefore, if (g, p) M-implements f then $g(\phi_{(g,p)}(u_i)) = f(u_i) = F$, and thus g assigns outcome F to *at least* one of the neighbors of u_i. Intuitively, the fact that the outcome function g assigns F to at least one between v_i and w_i corresponds to assigning "false" to literal x_i and $\bar{x}_i$.

We have one clause gadget for each clause; the gadget for clause C_j is depicted in Figure 2(b). Each clause C_j of Φ adds two new types to the domain D: c_j and d_j satisfying the following two non-contradicting inequalities:

$$c_j(F) - c_j(T) < d_j(F) - d_j(T), \tag{11}$$

$$d_j(T) - d_j(F) = \delta. \tag{12}$$

Node d_j has three edges directed towards the three variable gadgets corresponding to the variables appearing in the clause C_j. Specifically, if the clause C_j contains the literal x_i

then d_j is linked to the node v_i. Conversely, if C_j contains the literal $\bar{x}_i$ then d_j is connected to the node w_i. The social choice function f is defined by the labeling of the nodes; that is, $f(d_j) = T$ and $f(c_j) = F$.

Similarly to the variable gadget, we observe that (11) implies that c_j and d_j constitute a cycle of negative weight of length 2. Since $\phi_{(g,p)}(c_j) \in \{c_j, d_j\}$, then, by Lemma 13, it must be the case that $\phi_{(g,p)}(d_j) \notin \{c_j, d_j\}$. Since for any (g, p) that M-implements f it must be the case that g assigns T to d_j's best response, then g assigns outcome T to at least one of the neighbors of d_j from a variable gadget. We will see that this happens for all clauses if and only if the formula Φ is satisfiable. This concludes the description of the reduction.

We next prove that the reduction described above is correct. Suppose Φ is satisfiable, let τ be a satisfying assignment for Φ, let γ be a constant such that $\beta < \gamma < \delta$ and consider the following pair (g, p). For $i = 1, \cdots, n$, we set $g(a) = T$ and $p(a) = 0$ for all nodes a of the variable gadget for x_i except for v_i and w_i. Then, if $\tau(x_i) = 1$, we set $g(v_i) = T$, $p(v_i) = 0$, $g(w_i) = F$ and $p(w_i) = \gamma$. If instead $\tau(x_i) = 0$, we set $g(v_i) = F$, $p(v_i) = \gamma$, $g(w_i) = T$ and $p(w_i) = 0$. For $j = 1, \cdots, m$, we set $g(c_j) = g(d_j) = F$ and $p(c_j) = p(d_j) = 0$.

We now show that (g, p) M-implements f. We show this only for types u_i from variable gadgets and types d_j from clause gadgets, as for the other types the reasoning is immediate. Notice that by definition, g assigns F to exactly one of v_i and w_i and T to the other. Thus, denote by a the vertex $a \in \{v_i, w_i\}$ such that $g(a) = F$ and by b the vertex $b \in \{v_i, w_i\}$ such that $g(b) = T$. We show that a is u_i's best response under (g, p). Observe that $u_i(g(a))+p(a) = u_i(F)+\gamma > u_i(F) = u_i(g(t_i))+p(t_i)$. Therefore t_i is not u_i's best response. On the other hand, we have $u_i(g(b))+p(b) = u_i(T)$. But then, since $\gamma > \beta = u_i(T)-u_i(F)$, we have that a is u_i's best response under (g, p).

For d_j, we observe that, since τ satisfies clause C_j, there must exists at least one literal of C_j that is true under τ. By the definition of g, there exists at least one neighbor, call it a_j, of d_j from a variable gadget such that $g(a_j) = T$. We next show that a_j is d_j's best response. Notice that $p(a_j) = 0$. For all vertices b adjacent to d_j for which $g(b) = F$, we have $p(b) \leq \gamma$. But then, since $\gamma < \delta = d_j(T) - d_j(F)$ we have that a_j is d_j's best response under (g, p).

Conversely, consider an outcome function (g, p) that implements f and construct truth assignment τ as follows. Observe that, for any clause C_j, d_j and c_j constitute a cycle of negative weight and length 2. Moreover, c_j's best response is either c_j or d_j and thus, by Lemma 13, it must be the case that d_j's best response is a vertex, call it a_j, from a variable gadget such that $g(a_j) = T$. Then if $a_j = v_i$ for some i then we set $\tau(x_i) = 1$; if instead $a_j = w_i$ for some i we set $\tau(x_i) = 0$. Assignment τ (arbitrarily extended to unspecified variables) is easily seen to satisfy Φ.

The above discussion and the observation that the reduction can be carried out in polynomial time proves the following theorem.

Theorem 14 *The* Quasi-Linear Implementability *problem is NP-hard even for outcome sets of size* 2.

5 Conclusions

We have seen that is it NP-hard to decide if a given social choice function can be implemented even under the premise that the function does not admit a truthful implementation. Indeed, for these function it is NP-hard to decide if there is a non-truthful implementation, which in turn is the only way to implement them. An important factor here is the structure of the domain and the partial information, which we encode in the correspondence graph. In particular, we have the following results:

Correspondence Graph	No Payments	Payments and Quasi-linear Agent
Path	Polynomial [Th. 9]	Always implementable [3, Th. 4]
Directed acyclic	NP-hard [Th. 8]	Always implementable [3, Th. 4]
Arbitrary	NP-hard [Th. 8]	NP-hard [Th. 14]

Note that for directed acyclic graphs, the Quasi Linear Implementability Problem (where we ask implementability with payments) is trivially polynomial since all social choice functions are implementable wheras it is NP-hard to decide if an implementation without payments exists. So, it is also difficult to decide if payments are necessary or not for implementing a given function. Once again, this task becomes easy when restricting to truthful implementations [4].

Another interesting fact is that the problem without payments is difficult not because there are many possible outcomes, but because an agent may have several ways of misreporting his type. Indeed, the problem is easy if the agent has at most one way of lying (Theorem 9), but becomes NP-hard already for three (Theorem 8). The case of two remains open.

Finally, the fact that we consider the principal-agent model (the same as in [1]) only makes our negative results stronger since they obviously extend to the case of several agents (simply add extra agents whose corresponding function is $M(t) = \{t\}$).

On the other hand it remains open whether the positive result for graphs of outdegree at most 1 can be extended to many agents. Here the difficulty is the inter-dependence between the best response rules of the agents.

Acknowledgments. Work supported by EU through IP AEOLUS. The fourth author is also supported by DFG grant Kr 2332/1-2 within Emmy Noether Program. This paper is also published in the Proceedings of the First International Symposium on Algorithmic Game Theory (Burkhard Monien, Ulf-Peter Schroeder (Eds.), Lecture Notes in Computer Science 4997 Springer 2008, pp. 194-205).

References

[1] Jerry R. Green and Jean-Jacques Laffont. Partially Verifiable Information and Mechanism Design. *The Review of Economic Studies*, 53:447–456, 1986.

[2] Jean-Charles Rochet. A Condition for Rationalizability in a Quasi-Linear Context. *Journal of Mathematical Economics*, 16:191–200, 1987.

[3] Nirvikar Singh and Donald Wittman. Implementation with partial verification. *Review of Economic Design*, 6(1):63–84, 2001.

[4] Rakesh V. Vohra. Paths, cycles and mechanism design. Technical report, Kellogg School of Management, 2007.

Vincenzo Auletta
Dipartimento di Informatica ed Applicazioni,
Università di Salerno, Italy.
Email: `auletta@dia.unisa.it`

Paolo Penna
Dipartimento di Informatica ed Applicazioni,
Università di Salerno, Italy.
Email: `penna@dia.unisa.it`

Giuseppe Persiano
Dipartimento di Informatica ed Applicazioni,
Università di Salerno, Italy.
Email: giuper@dia.unisa.it

Carmine Ventre
Computer Science Department,
University of Liverpool, UK.
Email: Carmine.Ventre@liverpool.ac.uk

Complexity of comparison of influence of players in simple games

Haris Aziz[1]

Abstract

Coalitional voting games appear in different forms in multi-agent systems, social choice and threshold logic. In this paper, the complexity of comparison of influence between players in coalitional voting games is characterized. The possible representations of simple games considered are simple games represented by winning coalitions, minimal winning coalitions, weighted voting game or a multiple weighted voting game. The influence of players is gauged from the viewpoint of basic player types, desirability relations and classical power indices such as Shapley-Shubik index, Banzhaf index, Holler index, Deegan-Packel index and Chow parameters. Among other results, it is shown that for a simple game represented by minimal winning coalitions, although it is easy to verify whether a player has zero or one voting power, computing the Banzhaf value of the player is #P-complete. Moreover, it is proved that multiple weighted voting games are the only representations for which it is NP-hard to verify whether the game is linear or not. For a simple game with a set W^m of minimal winning coalitions and n players, a $O(n.|W^m| + n^2 log(n))$ algorithm is presented which returns 'no' if the game is non-linear and returns the strict desirability ordering otherwise. The complexity of transforming simple games into compact representations is also examined.

1 Introduction

1.1 Overview

Simple games are *yes/no* coalitional voting games which arise in various mathematical contexts. Simple games were first analysed by John von Neumann and Oskar Morgenstern in their monumental book *Theory of Games and Economic Behaviour* [25]. They also examined weighted voting games in which voters have corresponding voting weights and a coalition of voters wins if their total weights equal or exceed a specified quota. Neumann and Morgenstern [25] observe that minimal winning coalitions are a useful way to represent simple games. A similar approach has been taken in [11]. We examine the complexity of computing the influence of players in simple games represented by winning coalitions, minimal winning coalitions, weighted voting games and multiple weighted voting games.

1.2 Outline

In Section 2, we outline different representations and properties of simple games. In Section 3, compact representations of simple games are considered. After that, the complexity of computing the influence of players in simple games is considered from the point of view of player types (Section 4), desirability ordering (Section 5), power indices and Chow parameters (Section 6). The final Section includes a summary of results and some open problems.

[1]The author would like to thank Prof. Mike Paterson and anonymous referees for valuable suggestions.

2 Background

2.1 Definitions

Definitions 2.1. *A* simple voting game *is a pair* (N, v) *with* $v : 2^N \to \{0, 1\}$ *where* $v(\emptyset) = 0$, $v(N) = 1$ *and* $v(S) \leq v(T)$ *whenever* $S \subseteq T$. *A coalition* $S \subseteq N$ *is* winning *if* $v(S) = 1$ *and* losing *if* $v(S) = 0$. *A simple voting game can alternatively be defined as* (N, W) *where* W *is the set of winning coalitions. This is called the* extensive winning form. *A minimal winning coalition (MWC) of a simple game* v *is a winning coalition in which defection of any player makes the coalition losing. A set of minimal winning coalitions of a simple game* v *can be denoted by* $W^m(v)$. *A simple voting game can be defined as* (N, W^m). *This is called the* extensive minimal winning form.

For the sake of brevity, we will abuse the notation to sometimes refer to game (N, v) as v.

Lemma 2.2. *For a simple game* (N, W), W^m *can be computed in polynomial time.*

Proof. For any $S \in W$, remove elements from S until any further removals would make the coalition losing. The resultant coalition S' is a member of W^m. □

Definition 2.3. *A coalition* S *is* blocking *if its complement* $(N \setminus S)$ *is losing. For a simple game* $G = (N, W)$, *there is a* dual *game* $G^d = (N, W^d)$ *where* W^d *contains all the blocking coalitions in* G.

Definitions 2.4. *The simple voting game* (N, v) *where*
$W = \{X \subseteq N, \sum_{x \in X} w_x \geq q\}$ *is called a* weighted voting game*(WVG). A weighted voting game is denoted by* $[q; w_1, w_2, ..., w_n]$ *where* w_i *is the voting weight of player* i. *Usually,* $w_i \geq w_j$ *if* $i < j$.

Definitions 2.5. *An* m*-multiple weighted voting game (MWVG) is the simple game* $(N, v_1 \wedge \cdots \wedge v_m)$ *where the games* (N, v_t) *are the WVGs* $[q^t; w_1^t, \ldots, w_n^t]$ *for* $1 \leq t \leq m$. *Then* $v = v_1 \wedge \cdots \wedge v_m$ *is defined as:*

$$v(S) = \begin{cases} 1, & \text{if } v_t(S) = 1,\ \forall t,\ 1 \leq t \leq m. \\ 0, & \text{otherwise.} \end{cases}$$

The dimension of (N, v) *is the least* k *such that there exist WMGs* $(N, v_1), \ldots, (N, v_k)$ *such that* $(N, v) = (N, v_1) \wedge \ldots \wedge (N, v_k)$.

Definitions 2.6. *A WVG* $[q; w_1, \ldots, w_n]$ *is* homogeneous *if* $w(S) = q$ *for all* $S \in W^m$. *A simple game* (N, v) *is* homogeneous *if it can be represented by a homogeneous WVG. A simple game* (N, v) *is symmetric if* $v(S) = 1$, $T \subset N$ *and* $|S| = |T|$ *implies* $v(T) = 1$.

It is easy to see that symmetric games are homogeneous with a WVG representation of $[k; \underbrace{1, \ldots, 1}_{n}]$. That is the reason they are also called k-out-of-n simple games.

Banzhaf index [2] and Shapley-Shubik index [23] are two classic and popular indices to gauge the voting power of players in a simple game. They are used in the context of weighted voting games, but their general definition makes them applicable to any simple game.

Definition 2.7. *A player* i *is* critical *in a coalition* S *when* $S \in W$ *and* $S \setminus i \notin W$. *For each* $i \in N$, *we denote the number of coalitions in which* i *is critical in game* v *by the* Banzhaf value $\eta_i(v)$. *The* Banzhaf Index *of player* i *in weighted voting game* v *is* $\beta_i = \frac{\eta_i(v)}{\sum_{i \in N} \eta_i(v)}$.

Definitions 2.8. *The* Shapley-Shubik value *is the function κ that assigns to any simple game (N, v) and any voter i a value $\kappa_i(v)$ where $\kappa_i = \sum_{X \subseteq N}(|X|-1)!(n-|X|)!(v(X) - v(X - \{i\}))$. The* Shapley-Shubik index *of i is the function ϕ defined by $\phi_i = \frac{\kappa_i}{n!}$*

Definition 2.9. *([8]) For a simple game v, Chow parameters, CHOW(v) are $(|W_1|, \ldots |W_n|; |W|)$ where $W_i = \{S : S \subseteq N, i \in S\}$.*

2.2 Desirability relation and linear games

The individual desirability relations between players in a simple game date back at least to Maschler and Peleg [18].

Definitions 2.10. *In a simple game (N, v),*

- *A player i is* more desirable/influential *than player j ($i \succeq_D j$) if $v(S \cup \{j\}) = 1 \Rightarrow v(S \cup \{i\}) = 1$ for all $S \subseteq N \setminus \{i, j\}$.*
- *Players i and j are* equally desirable/influential *or* symmetric *($i \sim_D j$) if $v(S \cup \{j\}) = 1 \Leftrightarrow v(S \cup \{i\}) = 1$ for all $S \subseteq N \setminus \{i, j\}$.*
- *A player i is* strictly more desirable/influential *than player j ($i \succ_D j$) if i is more desirable than j, but if i and j are not equally desirable.*
- *A player i and j are* incomparable *if there exist $S, T \subseteq N \setminus \{i, j\}$ such that $v(S \cup \{i\}) = 1$, $v(S \cup \{j\}) = 0$, $v(T \cup \{i\}) = 0$ and $v(T \cup \{j\}) = 1$.*

Linear simple games are a natural class of simple games:

Definitions 2.11. *A simple game is* linear *whenever the desirability relation $\succeq_D$ is complete that is any two players i and j are comparable ($i \succ j$, $j \succ i$ or $i \sim j$).*

For linear games, the relation $R_\sim$ divides the set of voters N into equivalence classes $N/R_\sim = \{N_1, \ldots, N_t\}$ such that for any $i \in N_p$ and $j \in N_q$, $i \succ j$ if and only if $p < q$.

Definitions 2.12. *A simple game v is* swap robust *if an exchange of two players from two winning coalitions cannot render both losing. A simple game is* trade robust *if any arbitrary redistributions of players in a set of winning coalitions does not result in all coalitions becoming losing.*

It is easy to see that trade robustness implies swap robustness. Taylor and Zwicker [24] proved that a simple game can be represented by a WVG if and only if it is trade robust. Moreover they proved that a simple game being linear is equivalent to it being swap robust.

Taylor and Zwicker [24] show in Proposition 3.2.6 that v is linear if and only if $\succ_D$ is acyclic which is equivalent to $\succ_D$ being transitive. This is not guaranteed in other desirability relations defined over coalitions [9].

Proposition 2.13. *A simple game with three or fewer players is linear.*

Proof. For a game to be non-linear, we want to player 1 and 2 to be incomparable, i.e., there exist coalitions $S_1, S_2 \subseteq N \setminus \{1, 2\}$ such that $v(\{1\} \cup S_1) = 1$, $v(\{2\} \cup S_1) = 0$, $v(\{1\} \cup S_2) = 0$ and $v(\{2\} \cup S_2) = 1$. This is clearly not possible for $n = 1$ or 2. For $n = 3$, without loss of generality, v is non-linear only if $v(\{1\} \cup \emptyset) = 1$, $v(\{2\} \cup \emptyset) = 0$, $v(\{1\} \cup \{3\}) = 0$ and $v(\{2\} \cup \{3\}) = 1$. However the fact that $v(\{1\} \cup \emptyset) = 1$ and $v(\{1\} \cup \{3\}) = 0$ leads to a contradiction. □

3 Compact representations

Since WVGs and MWVG are compact representations of coalitional voting games, it is natural to ask which voting games can be represented by a WVG or MWVG and what is the complexity of answering the question. Deineko and Woeginger [6] show that it is NP-hard to verify the dimension of MWVGs. We know that every WVG is linear but not every linear game has a corresponding WVG. Carreras and Freixas,[3] show that there exists a six-player simple linear game which cannot be represented by a WVG. We now define problem *X-Realizable* as the problem to decide whether game v can be represented by form X.

Proposition 3.1. *WVG-Realizable is NP-hard for a MWVG.*

Proof. This follows directly from the proof by Deineko and Woeginger [6] that it is NP-hard to find the dimension of a MWVG. □

Proposition 3.2. *WVG-Realizable is in* P *for a simple game represented by its minimal winning, or winning, coalitions.*

This follows directly from Theorem 6 in [11]. The basic idea is that any simple game can be represented by linear inequalities. The idea dates back at least to [16] and the complexity of this problem was examined in the context of set covering problems. However it is one thing to know whether a simple game is WVG-Realizable and another thing to actually represent it by a WVG. It is not easy to represent a WVG-Realizable simple game by a WVG where all the weights are integers as the problem transforms from linear programming to integer programming.

Proposition 3.3. *(Follows from Theorem 1.7.4 of Taylor and Zwicker[11]) Any simple game is MWVG-Realizable.*

Taylor and Zwicker [24] showed that for every $n \geq 1$, there is simple game of dimension n. In fact it has been pointed out by Freixas and Puente [12] that that for every $n \geq 1$, there is linear simple game of dimension n. This shows that there is no clear relation between linearity and dimension of simple games. However it appears exceptionally hard to actually transform a simple game (N, W) or (N, W^m) to a corresponding MWVG. The dimension of a simple game may be exponential $(2^{(n/2)-1})$ in the number of players [24]. A simpler question is to examine the complexity of computing, or getting a bound for, the dimension of simple games.

4 Complexity of player types

A player in a simple game may be of various types depending on its level of influence.

Definitions 4.1. *For a simple game v on a set of players N, player i is a*

- dummy *if and only if* $\forall S \subseteq N$, *if* $v(S) = 1$, *then* $v(S \setminus \{i\}) = 1$;
- passer *if and only if* $\forall S \subseteq N$, *if* $i \in S$, *then* $v(S) = 1$;
- vetoer *if and only if* $\forall S \subseteq N$, *if* $i \notin S$, *then* $v(S) = 0$;
- dictator *if and only if* $\forall S \subseteq N$, $v(S) = 1$ *if and only if* $i \in S$.

It is easy to see that if a dictator exists, it is unique and all other players are dummies. This means that a dictator has voting power one, whereas all other players have zero voting power. We examine the complexity of identifying the dummy players in voting games. We already know that for the case of WVGs, Matsui and Matsui [19] proved that it is NP-hard to identify dummy players.

Lemma 4.2. *A player i in a simple game v is a dummy if and only if it is not present in any minimal winning coalition.*

Proof. Let us assume that player i is a dummy but is present in a minimal winning coalition. That means that it is critical in the minimal winning coalition which leads to a contradiction. Now let us assume that i is critical in at least one coalition S such that $v(S \cup \{i\}) = 1$ and $v(S) = 0$. In that case there is a $S' \subset S$ such that $S' \cup \{i\}$ is a MWC. □

Proposition 4.3. *For a simple game v,*

1. *Dummy players can be identified in linear time if v is of the form (N, W^m).*
2. *Dummy players can be identified in polynomial time if v is of the form (N, W).*

Proof. We examine each case separately:

1. By Lemma 4.2, a player is a dummy if and only if it is not in member of W^m
2. By Lemma 2.2, W^m can be computed in polynomial time.

□

From the definition, we know that a player has veto power if and only if the player is present in every winning coalition.

Proposition 4.4. *Vetoers can be identified in linear time for a simple game in the following representations: (N, W), (N, W^m), WVG and MWVG.*

Proof. We examine each of the cases separately:

1. (N, W): Initialize all players as vetoers. For each winning coalition, if a player is not present in the coalition, remove him from the list of vetoers.
2. (N, W^m): If there exists a winning coalition which does not contain player i, there will also exist a minimal winning coalition which does not contain i.
3. WVG: For each player i, i has veto power if and only if $w(N \setminus \{i\}) < q$.
4. MWVG: For each player i, i has veto power if and only if $N \setminus \{i\}$ is losing.

□

Proposition 4.5. *For a simple game represented by (N, W), (N, W^m), WVG or MWVG, it is easy to identify the passers and the dictator.*

Proof. We check both cases separately:

1. Passers: This follows from the definition of a passer. A player i is a passer if and only if $v(\{i\}) = 1$.
2. Dictator: It is easy to see that if a dictator exists in a simple game, it is unique. It follows from the definition of a dictator that a player i is a dictator in a simple game if $v(\{i\}) = 1$ and $v(N \setminus \{i\}) = 0$.

□

5 Complexity of desirability ordering

A *desirability ordering* on linear games is any ordering of players such that $1 \succeq_D 2 \succeq_D \ldots \succeq_D n$. A *strict* desirability ordering is the following ordering on players: $1 \circ 2 \circ \ldots \circ n$ where $\circ$ is either $\sim_D$ or $\succ_D$.

Proposition 5.1. *For a WVG:*

1. *A desirability ordering of players can be computed in polynomial time.*
2. *It is NP-hard to compute the strict desirability ordering of players.*

Proof. WVGs are linear games with a complete desirability ordering. For (1), it is easy to see that one desirability ordering of players in a WVG is the ordering of the weights. When $w_i = w_j$, then we know that $i \sim j$. Moreover, if $w_i > w_j$, then we know that i is at least as desirable as j, that is $i \succeq j$. For (2), the result immediately follows from the result by Matsui and Matsui [19] where they prove that it is NP-hard to check whether two players are symmetric. □

Let v be a MWVG of m WVGs on n players. It is easy to see that if there is an ordering of players such that such that $w_1^t \geq w_2^t \geq \ldots \geq w_n^t$ for all t, then v is linear. However, if an ordering like this does not exist, this does not imply that the game is not linear. The following is an example of a small non-linear MWVG:

Example 5.2. *In game $v = [10; 10, 9, 1, 0] \wedge [10; 9, 10, 0, 1]$, players 1 and 2 are incomparable. So, whereas simple games with 3 players are linear, it is easy to construct a 4 player non-linear MWVG.*

Proposition 5.3. *It is NP-hard to verify whether a MWVG is linear or not.*

Proof. We prove this by a reduction from an instance of the classical NP-hard PARTITION problem.

Name: PARTITION
Instance: A set of k integer weights $A = \{a_1, \ldots, a_k\}$.
Question: Is it possible to partition A, into two subsets $P_1 \subseteq A$, $P_2 \subseteq A$ so that $P_1 \cap P_2 = \emptyset$ and $P_1 \cup P_2 = A$ and $\sum_{a_i \in P_1} a_i = \sum_{a_i \in P_2} a_i$?

Given an instance of PARTITION $\{a_1, \ldots, a_k\}$, we may as well assume that $\sum_{i=1}^{k} a_i$ is an even integer, $2t$ say. We can transform the instance into the multiple weighted voting $v = v_1 \wedge v_2$ where $v_1 = [q; 20a_1, \ldots, 20a_k, 10, 9, 1, 0]$ and $v_2 = [q; 20a_1, \ldots, 20a_k, 9, 10, 0, 1]$ for $q = 10 + 20t$ and $k + 4$ is the number of players.

If A is a 'no' instance of PARTITION, then we see that a subset of weights $\{20a_1, \ldots, 20a_k\}$ cannot sum to $20t$. This implies that players $k + 1$, $k + 2$, $k + 3$, and $k + 4$ are not critical for any coalition. Since players $1, \ldots, k$ have the same desirability ordering in both v_1 and v_2, v is linear.

Now let us assume that A is a 'yes' instance of PARTITION with a partition (P_1, P_2). In that case players $k+1$, $k+2$, $k+3$, and $k+4$ are critical for certain coalitions. We see that $v(\{k+1\} \cup (\{k+4\} \cup P_1)) = 1$, $v(\{k+2\} \cup (\{k+4\} \cup P_1)) = 0$, $v(\{k+1\} \cup (\{k+3\} \cup P_1)) = 0$ and $v(\{k+2\} \cup (\{k+3\} \cup P_1)) = 1$. Therefore, players $k+1$ and $k+2$ are not comparable and v is not linear. □

Proposition 5.4. *For a simple game $v = (N, W^m)$, it can be verified in $O(n|W^m|)$ time if v is linear or not.*

Proof. Makino [17] proved that for a positive boolean function on n variables represented by the set of all minimal true vectors $minT(f)$, it can be checked in $O(n|minT(f)|)$ whether the function is *regular* (linear) or not. Makino's algorithm CHECK-FCB takes $minT(f)$ as input and outputs 'yes' if f is regular and 'no' otherwise. The proof involves encoding the minimal true vectors by a *fully condensed binary tree*. Then it follows that it can be verified in $O(n(|W^m|))$ whether a simple game $v = (N, W^m)$ is linear or not. □

Corollary 5.5. *For a simple game $v = (N, W)$, it can be verified in polynomial time if v is linear or not.*

Proof. We showed earlier that (N, W) can be transformed into (N, W^m) in polynomial time. After that we can use Makino's method [17] to verify whether the game is linear or not. □

Muroga [20] cites Winder [26] for a result concerning comparison between boolean variables and their incidence in prime implicants of a boolean function. Hilliard [14] points out that this result can be used to check the desirability relation between players in WVG-Realizable simple games. We generalize Winder's result by proving both sides of the implications and extend Hilliard's observation to that of linear simple games.

Proposition 5.6. *Let $v = (N, W^m)$ be a linear simple game and let $d_{k,i} = |\{S : i \in S, S \in W^m, |S| = k\}|$. Then for two players i and j,*

1. *$i \sim_D j$ if and only if $d_{k,i} = d_{k,j}$ for $k = 1, \ldots n$.*
2. *$i \succ_D j$ if and only if for the smallest k where $d_{k,i} \neq d_{k,j}$, $d_{k,i} > d_{k,j}$.*

Proof. 1. ($\Rightarrow$) Let us assume $i \sim_D j$. Then by definition, $v(S \cup \{j\}) = 1 \Leftrightarrow v(S \cup \{i\}) = 1$ for all $S \subseteq N \setminus \{i, j\}$. So $S \cup \{i\} \in W^m$ if and only if $S \cup \{j\} \in W^m$. Therefore, $d_{k,i} = d_{k,j}$ for $k = 1, \ldots n$.

($\Leftarrow$) Let us assume that $i \nsim_D j$. Since v is linear, i and j are comparable. Without loss of generality, we assume that $i \succ_D j$. Then there exists a coalition $S \setminus \{i, j\}$ such that $v(S \cup \{i\}) = 1$ and $v(S \cup \{j\}) = 0$ and suppose $|S| = k - 1$. If $S \cup \{i\} \in W^m$, then $d_{k,i} > d_{k,j}$. If $S \cup \{i\} \notin W^m$ then there exists $S' \subset S$ such that $S' \cup \{i\} \in W^m$. Thus there exists $k' < k$ such that $d_{k',i} > d_{k',j}$.

2. ($\Rightarrow$) Let us assume that $i \succ_D j$ and let k' be the smallest integer where $d_{k',i} \neq d_{k',j}$. If $d_{k',i} < d_{k',j}$, then there exists a coalition S such that $S \cup \{j\} \in W^m$, $S \cup \{i\} \notin W^m$ and $|S| = k' - 1$. $S \cup \{i\} \notin W^m$ in only two cases. The first possibility is that $v(S \cup \{i\}) = 0$, but this is not true since $i \succ_D j$. The second possibility is that there exists a coalition $S' \subset S$ such that $S' \cup \{i\} \in W^m$. But that would mean that $v(S' \cup \{i\}) = 1$ and $v(S' \cup \{j\}) = 0$. This also leads to a contradiction since k' is the smallest integer where $d_{k',i} \neq d_{k',j}$.

   ($\Leftarrow$) Let us assume that for the smallest k where $d_{k,i} \neq d_{k,j}$, $d_{k,i} > d_{k,j}$. This means there exists a coalition S such that $S \cup \{i\} \in W^m$, $S \cup \{j\} \notin W^m$ and $|S| = k - 1$. This means that either $v(S \cup \{j\}) = 0$ or there exists a coalition $S' \subset S$ such that $S' \cup \{i\} \in W^m$. If $v(S \cup \{j\}) = 0$, that means $i \succ_D j$. If there exists a coalition $S' \subset S$ such that $S' \cup \{j\} \in W^m$, then $d_{k',j} > d_{k',i}$ for some $k' < k$. This leads to a contradiction.

□

We can use this theorem and Makino's 'CHECK-FCB' algorithm [17] to make an algorithm which takes as input a simple game (N, W^m) and returns NO if the game is not linear and returns the strict desirability ordering otherwise.

Algorithm 1 Strict-desirability-ordering-of-simple-game

Input: Simple game $v = (N, W^m)$ where $N = \{1, \ldots, n\}$ and $W^m(v) = \{S_1, \ldots, S_{|W^m|}\}$.
Output: NO if v is not linear. Otherwise output desirability equivalence classes starting from most desirable, if, v is linear.

$X =$ CHECK-FCB(W^m)
if $X = NO$ **then**
  return NO
else
  Initialize an $n \times n$ matrix D where entries $d_{i,j} = 0$ for all i and j in N
  for $i = 1$ to $|W^m|$ **do**
    for each player x in S_i **do**
      $d_{|S_i|,x} \leftarrow d_{|S_i|,x} + 1$
    end for
  end for
  return $\mathsf{classify}(N, D, 1)$
end if

Algorithm 2 classify

Input: set of integers $\mathsf{class_{index}}$, $n \times n$ matrix D, integer k.
Output: subclasses.

if $k = n + 1$ or $|\mathsf{class_{index}}| = 1$ **then**
  return $\mathsf{class_{index}}$
end if
$s \leftarrow |\mathsf{class_{index}}|$
$\mathsf{mergeSort}(\mathsf{class_{index}})$ in descending order such that $i > j$ if $d_{k,i} > d_{k,j}$.
for $i = 2$ to s **do**
  $\mathsf{subindex} \leftarrow 1$; $\mathsf{class_{index.subindex}} \leftarrow \mathsf{class_{index}}[1]$
  if $d_{k,\mathsf{class_{index}}[i]} = d_{k,\mathsf{class_{index}}[i-1]}$ **then**
    $\mathsf{class_{index.subindex}} \leftarrow \mathsf{class_{index.subindex}} \cup \mathsf{class_{index}}[i]$
  else if $d_{k,\mathsf{class_{index}}[i]} < d_{k,\mathsf{class_{index}}[i-1]}$ **then**
    $\mathsf{subindex} \leftarrow \mathsf{subindex} + 1$
    $\mathsf{class_{index.subindex}} \leftarrow \{\mathsf{class_{index}}[i]\}$
  end if
end for
$\mathsf{Returnset} \leftarrow \emptyset$
$A \leftarrow \emptyset$
for $j = 1$ to $\mathsf{subindex}$ **do**
  $A \leftarrow \mathsf{classify}(\mathsf{class_{index.j}}, D, k + 1)$
  $\mathsf{Returnset} \leftarrow A \cup \mathsf{Returnset}$
end for
return $\mathsf{Returnset}$

Proposition 5.7. *The time complexity of Algorithm 1 is* $O(n.|W^m| + n^2 log(n))$

Proof. The time complexity of $CHECK - FCB$ is $O(n.|W^m|)$. The time complexity of computing matrix D is $O(\mathsf{Max}(|W^m|, n^2)$. For each iteration, sorting of sublists requires at most $O(nlog(n))$ time. There are at most n loops. Therefore the total time complexity is $O(n.|W^m|) + O(\mathsf{Max}(|W^m|, n^2) + O(n^2 log(n)) = O(n.|W^m| + n^2 log(n))$.

□

Corollary 5.8. *The strict desirability ordering of players in a linear simple game $v = (N, W)$ can be computed in polynomial time.*

Proof. The proof follows directly from the Algorithm. Moreover, we know that the set of all winning coalitions can be transformed into a set of minimal winning coalitions in polynomial time. □

6 Power indices and Chow parameters

Apart from the Banzhaf and Shapley-Shubik indices, there are other indices which are also used. Both the Deegan-Packel index [5] and the Holler index [15] are based on the notion of minimal winning coalitions. Minimal winning coalitions are significant with respect to coalition formation [4]. The Holler index, H_i of a player i in a simple game corresponds to the Banzhaf index with one difference: only swings in minimal winning coalitions contribute towards the Holler index.

Definitions 6.1. *We define the Holler value M_i as $\{S \in W^m : i \in S\}$. The Holler index which is called the* public good index *is defined by $H_i(v) = \frac{|M_i|}{\sum_{j \in N} |M_j|}$. The Deegan Packel index for player i in voting game v is defined by $D_i(v) = \frac{1}{|W^m|} \sum_{S \in M_i} \frac{1}{|S|}$.*

Compared to the Banzhaf index and the Shapley-Shubik index, both the Holler index and the Deegan-Packel index do not always satisfy the monotonicity condition. In [19], Matsui and Matsui prove that it is NP-hard to compute the Banzhaf index, Shapley-Shubik index and Deegan-Packel index of a player. We can use a similar technique to also prove that it is NP-hard to compute the Holler index of players in a WVG. This follows directly from the fact that it is NP-hard to decide whether a player is dummy or not. Prasad and Kelly [21] and Deng and Papadimitriou [7] proved that for WVGs, computing the Banzhaf values and Shapley-Shubik values is #P-parsimonious-complete and #P-metric-complete respectively. (For details on #P-completeness and associated reductions, see [10]). Unless specified, reductions considered with #P-completeness will be Cook reductions (or polynomial-time Turing reductions).

What we see is that although it is NP-hard to compute the Holler index and Deegan-Packel of players in a WVG, the Holler index and Deegan-Packel of players in a simple game represented by its MWCs can be computed in linear time:

Proposition 6.2. *For a simple game (N, W^m), the Holler index and Deegan-Packel index for all players can be computed in linear time.*

Proof. We examine each of the cases separately:

- Initialize M_i to zero. Then for each $S \in W^m$, if $i \in S$, increment M_i by one.
- Initialize d_i to zero. Then for each $S \in W^m$, if $i \in S$, increment d_i, by $\frac{1}{|S|}$. Then $D_i = \frac{d_i}{|W^m|}$.

□

Proposition 6.3. *For a simple game $v = (N, W)$, the Banzhaf index, Shapley Shubik index, Holler index and Deegan-Packel index can be computed in polynomial time.*

Proof. The proof follows from the definitions. We examine each of the cases separately:

- Holler index: Transform W into W^m and then compute the Holler indices.

- Deegan-Packel: Transform W into W^m and then compute the Deegan-Packel indices.
- Banzhaf index: Initialize Banzhaf values of all players to zero. For each $S \in W$, check if the removal of a player results in S becoming losing (not a member of W). In that case increment the Banzhaf value of that player by one.
- Shapley-Shubik index: Initialize Shapley values of all players to zero. For each $S \in W$, check if the removal of a player results in S becoming losing (not a member of W). In that case increment the Shapley value of the player by $(|S|-1)!(n-|S|)!$.

The time complexity for all cases is polynomial in the order of the input. □

For a simple game (N, W^m), listing W the winning coalitions may take time exponential in the number of players. For example, let there be only one minimal winning coalition S which contains players $1, \ldots, \lceil n/2 \rceil$. Then the number of winning coalitions to list is exponential in the number of players. Moreover, if $|W^m| > 1$, minimal winning coalitions can have common supersets. It is shown below that for a simple game (N, W^m), even counting the total number of winning coalitions is #P-complete. Moreover, whereas it is polynomial time easy to check if a player has zero voting power (a dummy) or whether it has voting power 1 (dictator), it is #P-complete to find the actual Banzhaf or Shapley-Shubik index of the player.

Proposition 6.4. *For a simple game $v = (N, W^m)$, the problem of computing the Banzhaf values of players is #P-complete.*

Proof. The problem is clearly in #P. We prove the #P-hardness of the problem by providing a reduction from the problem of computing $|W|$. Ball and Provan [1] proved that computing $|W|$ is #P-complete. Their proof is in context of reliability functions so we first give the proof in terms of simple games. It is known that known [22] that counting the number of vertex covers is #P-complete (a vertex cover in a graph $G = (V, E)$ is a subset C of V such that every edge in E has at least one endpoint in C). Now take a simple game $v = (N, W^m)$ where for any $S \in W^m$, $|S| = 2$. Game v has a one-to-one correspondence with a graph $G = (V, E)$ such that $N = V$ and $\{i, j\} \in W^m$ if and only if $\{i, j\} \in E(G)$. In that case the total number of losing coalitions in v is equal to the number of vertex covers of G. Therefore the total number of winning coalitions is equal to $2^n -$(number of vertex covers of G) and computing $|W|$ is #P-complete.

Now we take a game $v = (N, W^m)$ and convert it into another game $v' = (N \cup \{n+1\}, W^m(v'))$ where for each $S \in W^m(v)$, $S \cup \{n+1\} \in W^m(v')$. In that case computing $|W(v)|$ is equivalent to computing the Banzhaf value of player $n+1$ in game v'. Therefore, computing Banzhaf values of players in games represented by MWCs is #P-hard. □

It follows from the proof that computing *power of collectivity to act*$(\frac{|W|}{2^n})$ and Chow parameters for a simple game (N, W^m) is #P-complete. Goldberg remarks in the conclusion of [13] that computing the Chow parameters of a WVG is #P-complete. It is easy to prove this. The problem of computing $|W|$ and $|W_i|$ for any player i is in #P since a winning coalition can be verified in polynomial time. It is easy to reduce in polynomial time the counting version of the SUBSET-SUM problem to counting the number of winning coalitions. Moreover, for any WVG $v = [q; w_1, \ldots, w_n]$, $|W(v)|$ is equal to $|W_{n+1}(v')|$ where v' is $[q; w_1, \ldots, w_n, 0]$. Therefore computing $|W_i|$ and $|W|$ for a WVG is #P-complete.

7 Conclusion

A summary of results has been listed in Table 1. A question mark indicates that the specified problem is still open. It is conjectured that computing Shapley values is #P-complete and

Table 1: Summary of results

	(N, W)	(N, W^m)	WVG	MWVG
IDENTIFY-DUMMIES	P	linear	NP-hard	NP-hard
IDENTIFY-VETOERS	linear	linear	linear	linear
IDENTIFY-PASSERS	linear	linear	linear	linear
IDENTIFY-DICTATOR	linear	linear	linear	linear
CHOW PARAMETERS	linear	#P-complete	#P-complete	#P-complete
IS-LINEAR	P	P	(Always linear)	NP-hard
DESIRABILITY-ORDERING	P	P	P	NP-hard
STRICT-DESIRABILITY	P	P	NP-hard	NP-hard
BANZHAF-VALUES	P	#P-complete	#P-complete	#P-complete
BANZHAF-INDICES	P	?	NP-hard	NP-hard
SHAPLEY-SHUBIK-VALUES	P	?	#P-complete	#P-complete
SHAPLEY-SHUBIK-INDICES	P	?	NP-hard	NP-hard
HOLLER-INDICES	P	linear	NP-hard	NP-hard
DEEGAN-PACKEL-INDICES	P	linear	NP-hard	NP-hard

it is NP-hard to compute Banzhaf indices for a simple game represented by (N, W^m). It is found that although WVG, MWVG and even (N, W^m) is a relatively compact representation of simple games, some of the important information encoded in these representations can apparently only be accessed by unraveling these representations. There is a need for a greater examination of transformations of simple games into compact representations.

References

[1] M. O. Ball and J. S. Provan. Disjoint products and efficient computation of reliability. *Oper. Res.*, 36(5):703–715, 1988.

[2] J. F. Banzhaf. Weighted voting doesn't work. *Rutgers Law Review*, 19:317–343, 1965.

[3] F. Carreras and J. Freixas. Complete simple games. *Mathematical Social Sciences*, 32(2):139–155, October 1996.

[4] B. B. de Mesquita. Minimum winning coalition, in politics. *Neil J. Smelser and Paul B. Baltes, Editor(s)-in-Chief, International Encyclopedia of the Social & Behavioral Sciences*, pages 9889–9891, 2001.

[5] J. Deegan and E. Packel. A new index of power for simple n-person games. *International Journal of Game Theory*, 7(2):113123, 1978.

[6] V. G. Deineko and G. J. Woeginger. On the dimension of simple monotonic games. *European Journal of Operational Research*, 170(1):315–318, 2006.

[7] X. Deng and C. H. Papadimitriou. On the complexity of cooperative solution concepts. *Math. Oper. Res.*, 19(2):257–266, 1994.

[8] P. Dubey and L. S. Shapley. Mathematical properties of the banzhaf power index. *Mathematics of Operations Research*, 4(2):99–131, 1979.

[9] E. Einy. The desirability relation of simple games. *Mathematical Social Sciences*, 10(2):155–168, October 1985.

[10] P. Faliszewski and L. A. Hemaspaandra. The complexity of power-index comparison. In *AAIM*, pages 177–187, 2008.

[11] J. Freixas, X. Molinero, M. Olsen, and M. Serna. The complexity of testing properties of simple games. *ArXiv e-prints*, 2008.

[12] J. Freixas and M. A. Puente. Dimension of complete simple games with minimum. *European Journal of Operational Research*, 127(2):555–568, July 2008.

[13] P. W. Goldberg. A bound on the precision required to estimate a boolean perceptron from its average satisfying assignment. *SIAM J. Discret. Math.*, 20(2):328–343, 2006.

[14] M. Hilliard. *Weighted voting theory and applications.* Tech. Report No. 609, chool of Operations Research and Industrial Engineering, Cornell University, 1983.

[15] M. Holler. Forming coalitions and measuring voting power. *Political Studies*, 30(2):262271, 1982.

[16] S.-T. Hu. Threshold logic. *University of California Press, Berkeley and Los Angeles*, 1965.

[17] K. Makino. A linear time algorithm for recognizing regular boolean functions. *J. Algorithms*, 43(2):155–176, 2002.

[18] M. Maschler and B. Peleg. A characterization, existence proof and dimension bounds for the kernel of a game. *Pacific J. Math*, 18(2):289–328., 1966.

[19] T. Matsui and Y. Matsui. A survey of algorithms for calculating power indices of weighted majority games. *Journal of the Operations Research Society of Japan*, 43(7186), 2000.

[20] S. Muroga. *Threshold logic and Its Applications.* Wiley Interscience, New York, 1971.

[21] K. Prasad and J. S. Kelly. NP-completeness of some problems concerning voting games. *Int. J. Game Theory*, 19(1):1–9, 1990.

[22] J. S. Provan and M. O. Ball. The complexity of counting cuts and of computing the probability that a graph is connected. *SIAM J. Comput.*, 12(4):777–788, 1983.

[23] L. S. Shapley. A value for n person games. *A. E. Roth,editor, The Shapley value*, page 3140, 1988.

[24] A. Taylor and W. Zwicker. *Simple Games: Desirability Relations, Trading, Pseudoweightings.* Princeton University Press, New Jersey, first edition, 1999.

[25] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior.* Princeton University Press, 1944.

[26] R. Winder. *Threshold logic, Ph.D. Thesis.* Mathematics Department, Princeton University, New Jersey, 1962.

Haris Aziz
Computer Science Department, University of Warwick
Coventry, UK, CV4 7AL
Email: `haris.aziz@warwick.ac.uk`

Divide and Conquer: False-Name Manipulations in Weighted Voting Games

Yoram Bachrach and Edith Elkind

Abstract

Weighted voting is a well-known model of cooperation among agents in decision-making domains. In such games, each player has a weight, and a coalition of players wins if its total weight meets or exceeds a given quota. Usually, the agents' power in such games is measured by a *power index*, such as, e.g., Shapley–Shubik index. In this paper, we study how an agent can manipulate its voting power (as measured by Shapley–Shubik index) by distributing his weight among several false identities. We show that such manipulations can indeed increase an agent's power and provide upper and lower bounds on the effects of such manipulations. We then study this issue from the computational perspective, and show that checking whether a beneficial split exists is NP-hard. We also discuss efficient algorithms for restricted cases of this problem, as well as randomized algorithms for the general case.

1 Introduction

Collaboration and cooperative decision-making are important issues in many types of interactions among self-interested agents. In many situations, agents must take a joint decision leading to a certain outcome, which may have a different impact on each of the agents. A standard and well-studied way of doing so is by means of voting, and in recent years, there has been a lot of research on applications of voting to multiagent systems as well as on computational aspects of various voting procedures. One of the key issues in this domain is how to measure the *power* of each voter, i.e., his impact on the final outcome. In particular, this question becomes important when the agents have to decide how to distribute the payoffs resulting from their joint action: a natural approach would be to pay each agent according to his contribution, i.e., his voting power.

This issue is traditionally studied within the framework of *weighted voting games*, which provide a model of decision-making in many political and legislative bodies, and have also been applied in the context of multiagent systems. In such a game, each of the agents has a weight, and a coalition of agents wins the game if the sum of the weights of its participants exceeds a certain quota. Having a larger weight makes it easier for an agent to affect the outcome; however, the agent's power is not always proportional to his weight. For example, if the quota is so high that the only winning coalition is the one that includes all agents, intuitively, all agents have equal power, irrespective of their weight. This idea is formalized using the concept of a *power index*, which is a systematic way of measuring a player's influence in a weighted voting game. There are several ways to define power indices. One of the most popular approaches relies on the fact that weighted voting games form a subclass of *coalitional games*, and therefore one can use the terminology and solution concepts that have been developed in the context of general coalitional games. In particular, an important notion in coalitional games is that of the Shapley value [12], which is a classical way to distribute the gains of the grand coalition in general coalitional games. In the context of weighted voting, the Shapley value (also known as the Shapley–Shubik power index [13]) provides a convenient measure of an agent's power and has been widely studied from both a normative and a computational perspective.

As suggested above, power indices measure the agents' power and can be used to de-

termine their payoffs. However, to be applicable in real-world scenarios, this approach has to be resistant to dishonest behavior, or manipulation, by the participating agents. In this paper we study the effects of one form of manipulation in weighted voting games, namely, false-name voting. Under this manipulation, an agent splits his weight between himself and a "fake" agent who enters the game. While the total weight of all identities of the cheating agent remains the same, his power (as measured by the Shapley–Shubik power index) may change. In open anonymous environments, such as the internet, this behavior is virtually impossible to detect, and therefore it presents a challenge to the designers of multiagent systems that rely on weighted voting. The goal of this paper is to measure the effects of false-name voting and analyze its computational feasibility. Our main results here are as follows:

- We precisely quantify the worst-case effect of false-name voting on agents' payoffs. Namely, we show that in an n-player game, false-name voting can increase an agent's payoff by a factor of $2n/(n+1)$, and this bound is tight. On the other hand, we show that false-name voting can decrease an agent's payoff by a factor of $(n+1)/2$, and this bound is also tight.
- We demonstrate that finding a successful manipulation is not a trivial task by proving that it is NP-hard to verify if a beneficial split exists. However, we show that if all weights are polynomially bounded, the problem can be solved in polynomial time, and discuss efficient randomized algorithms for this problem.

We also study several variants of the problem, such as splitting into more than two identities, as well as the dual problem of manipulation by merging, where several agents pretend to be one.

2 Related Work

In his seminal paper, Shapley [12] considered coalitional games and the question of fair allocation of the utility gained by the grand coalition. The solution concept introduced in this paper became known as the *Shapley value* of the game. The subsequent paper [13] studies the Shapley value in the context of simple coalitional games, where it is usually referred to as the *Shapley–Shubik power index*. Another measure of a player's influence in voting games is the Banzhaf power index [1].

Both of these power indices have been well studied. Straffin [14] shows that each index reflects certain conditions in a voting body. Paper [5] describes certain axioms that characterize these two indices, as well as several others. These indices were used to analyze the voting structures of the European Union Council of Ministers and the IMF [7, 6].

The applicability of the power indices to measuring political power in various domains has raised the question of finding tractable ways to compute them. However, this problem appears to be computationally hard. Indeed, the naive algorithm for calculating the Shapley value (or the Shapley–Shubik power index) considers all permutations of the players and hence runs in exponential time. Moreover, paper [3] shows that computing the Shapley value in weighted voting games is #P-complete.

Despite this hardness result, several works show how to compute these power indices in some *restricted domains*, or discuss ways to *approximate* them [9, 11, 8, 4]. A good survey of algorithms for calculating power indices in weighted voting games is [10]. Many of these approaches work well in practice, which justifies the use of these indices as payoff distribution schemes in multiagent domains.

False-name manipulation has been studied in the context of non-cooperative games such as auctions [16, 17], and, more recently, also in cooperative games [2, 15]. However, to the

best of our knowledge, ours is the first paper to systematically study this type of behavior in weighted voting.

3 Preliminaries and Notation

Coalitional Games A *coalitional game* $G = (I, v)$ is given by a set of agents $I = \{a_1, \ldots, a_n\}$, $|I| = n$, and a function $v : 2^I \to \mathbb{R}$ that maps any subset (coalition) of the agents to a real value. This value is the total utility these agents can guarantee to themselves when working together. To simplify notation, we will sometimes write i instead of a_i.

A coalitional game is *simple* if v can only take values 0 and 1, i.e., $v : 2^I \to \{0, 1\}$. In such games, we say that a coalition $C \subseteq I$ *wins* if $v(C) = 1$, and *loses* if $v(C) = 0$. An agent i is *critical*, or *pivotal*, to a winning coalition C if the agent's removal from that coalition would make it a losing coalition: $v(C) = 1$, $v(C \setminus \{i\}) = 0$.

Weighted Voting Games A *weighted voting game* G is a simple game that is described by a vector of players' *weights* $\mathbf{w} = (w_1, \ldots, w_n)$ and a *quota* q. We write $G = [w_1, \ldots, w_n; q]$, or $G = [\mathbf{w}; q]$. In these games, a coalition is winning if its total weight meets or exceeds the quota. Formally, for any $J \subseteq I$ we have $v(J) = 1$ if $\sum_{i \in J} w_i \geq q$ and $v(J) = 0$ otherwise. We will often write $w(J)$ to denote the total weight of a coalition J, i.e., $w(J) = \sum_{i \in J} w_i$. Also, we set $w_{\max} = \max_{i=1,\ldots,n} w_i$.

Shapley Value Intuitively, the Shapley value of an agent is determined by his marginal contribution to possible coalitions. Let Π_n be the set of all possible permutations (orderings) of n agents. Each $\pi \in \Pi_n$ is a one-to-one mapping from $\{1, \ldots, n\}$ to $\{1, \ldots, n\}$. Denote by $S_\pi(i)$ the predecessors of agent i in π, i.e., $S_\pi(i) = \{j \mid \pi(j) < \pi(i)\}$. The Shapley value of the ith agent in a game $G = (I, v)$ is denoted by $\varphi_G(i)$ and is given by the following expression:

$$\varphi_G(i) = \frac{1}{n!} \sum_{\pi \in \Pi_n} [v(S_\pi(i) \cup \{i\}) - v(S_\pi(i))]. \tag{1}$$

We will occasionally abuse notation and say that an agent i is pivotal for a permutation π if it is pivotal for the coalition $S_\pi(i) \cup \{i\}$.

The Shapley-Shubik power index is simply the Shapley value in a simple coalitional game (and therefore in the rest of the paper we will use these terms interchangeably). In such games the value of a coalition is either 0 or 1, so the formula (1) simply counts the fraction of all orderings of the agents in which agent i is critical for the coalition formed by his predecessors and himself. The Shapley-Shubik power index thus reflects the assumption that when forming a coalition, any *ordering* of the agents entering the coalition has an equal probability of occurring, and expresses the probability that agent i is critical.

While there exist several other approaches to determining the players' influence in a game, the Shapley value has many useful properties that make it very convenient to work with. We will make use of two of these properties, namely, the normalization property and the dummy player property. The former simply states that the sum of Shapley values of all players is equal to 1. The latter claims that the value of a dummy player is 0, where a player i is called a *dummy* if he contributes nothing to any coalition, i.e., for any $C \subseteq I$ we have $v(C \cup \{i\}) = v(C)$. It is easy to verify from the definitions that Shapley value has both of these properties.

4 False-Name Manipulations

As discussed in the introduction, it might be possible for a player to change his total payoff by splitting his weight between several identities. We will start by providing a few examples of such scenarios.

Example 1. *Consider a voting game $G = [8, 8, 1, 2; 11]$, i.e., a game with a quota of $q = 11$, and four agents $a_1, \ldots, a_4$, where a_1 has weight $w_1 = 8$, a_2 has weight $w_2 = 8$, a_3 has weight $w_3 = 1$, and agent a_4 has weight $w_4 = 2$.*

Using formula (1) to compute the Shapley value of a_4, we get $\varphi_G(a_4) = 4/24$. Now, suppose agent a_4 splits his weight equally between two new identities a'_4 and a''_4, resulting in a new game $G' = [8, 8, 1, 1, 1; 11]$. Calculating the Shapley value of a'_4 and a''_4 in this game, we get $\varphi_{G'}(a'_4) = \varphi_{G'}(a''_4) = 12/120$. Hence, although the total weight *of all identities of player a_4 is the same as in the original game, the total* power *held by the agent and his false-name "accomplice" has increased, since $\varphi_{G'}(a'_4) + \varphi_{G'}(a''_4) = 24/120 > \varphi_G(a_4) = 4/24$.*

Our next example shows that false-name manipulations are not necessarily beneficial.

Example 2. *Consider a game $G = [3, 3, 2; 4]$. Agent a_3 with the weight of $w_3 = 2$ has Shapley value of $\varphi_G(a_3) = 2/6$. Splitting his weight between two identities a'_3 and a''_3 results in a game $G' = [3, 3, 1, 1; 4]$. The Shapley values of the two identities a'_3 and a''_3 of the agent a_3 in the new game are $\varphi_{G'}(a'_3) = \varphi_{G'}(a''_3) = 4/24$, so we have $\varphi_G(a_3) = \varphi_{G'}(a'_3) + \varphi_{G'}(a''_3)$. Consequently, a_3 neither gained nor lost power by splitting.*

Moreover, weight-splitting can be risky for the manipulator, as illustrated by the following example.

Example 3. *Consider a game $G = [2, 2, 2; 5]$. Agent a_3 with the weight of $w_3 = 2$ has Shapley value of $\varphi_G(a_3) = 2/6$. By splitting into two agents, each with a weight of 1, this agent can get the following game $G' = [2, 2, 1, 1; 5]$. The Shapley values of the two agents a'_3 and a''_3 in the new game are $\varphi_{G'}(a'_3) = \varphi_{G'}(a''_3) = 2/24$, so we have $\varphi_G(a_3) > \varphi_{G'}(a'_3) + \varphi_{G'}(a''_3)$. Hence, the splitting agent has lost power by splitting, i.e., it was harmful for him to split.*

4.1 Effects of Manipulation: Upper and Lower Bounds

We have seen that an agent can both increase and decrease his total payoff by splitting his weight. In this subsection, we provide upper and lower bounds on how much he can change his payoff by doing so. We restrict our attention to the case of splitting into two identities; the general case is briefly discussed in Section 7.

To simplify notation, in the rest of this section we assume that in the original game $G = [w_1, \ldots, w_n; q]$ the manipulator is agent a_n, and he splits into two new identities a'_n and a''_n, resulting in a new game G'.

Theorem 4. *For any game $G = [w_1, \ldots, w_n; q]$ and any split of a_n's into a'_n and a''_n, we have $\varphi_{G'}(a'_n) + \varphi_{G'}(a''_n) \leq \frac{2n}{n+1}\varphi_G(a_n)$, i.e., the manipulator cannot gain more than a factor of $2n/(n+1) < 2$ by splitting his weight between two identities. Moreover, this bound is tight, i.e., there exists a game in which agent a_n increases his payoff by a factor of $2n/(n+1)$ by splitting into two identities.*

Proof. Fix a split of a_n into a'_n and a''_n. Let Π_{n-1} be the set of all permutations of the first $n-1$ agents. Consider any $\pi \in \Pi_{n-1}$. Let $P(\pi)$ be the set of all permutations of the agents in G' that can be obtained by inserting a'_n and a''_n into π. Let Π^*_{n+1} be the set of all permutations π^* of agents in G' such that a'_n or a''_n is pivotal for π^*. Finally, Let $P^*(\pi, k)$

be the subset of $P(\pi) \cap \Pi^*_{n+1}$ that consists of all permutations $\pi' \in P(\pi)$ in which at least one of the a'_n and a''_n appears between the kth and the $(k+1)$st element of π' and is pivotal for π'. Every permutation in Π^*_{n+1} appears in one of the sets $P^*(\pi, k)$ for some π, k, so we have

$$\varphi_{G'}(a'_n) + \varphi_{G'}(a''_n) = \frac{|\Pi^*_{n+1}|}{(n+1)!} \leq \frac{1}{(n+1)!} \sum_{\pi,k} |P^*(\pi, k)|.$$

On the other hand, it is not hard to see that $|P^*(\pi, k)| \leq 2n$ for any π, k: there are two ways to place a'_n and a''_n between the kth and the $(k+1)$st element of π, $n-1$ permutations in $P^*(\pi, k)$ in which a'_n appears after the kth element of π, but a''_n is not adjacent to it, and $n-1$ permutations in $P^*(\pi, k)$ in which a''_n appears after the kth element of π, but a'_n is not adjacent to it. Moreover, if $P^*(\pi, k)$ is not empty, then a_n is pivotal for the permutation $f(\pi, k)$ obtained from π by inserting a_n after the kth element of π. Moreover, if $(\pi_1, k_1) \neq (\pi_2, k_2)$ then $f(\pi_1, k_1) \neq f(\pi_2, k_2)$. Hence,

$$\varphi_G(a_n) \geq \frac{1}{n!} \sum_{\pi,k:P^*(\pi,k)\neq\emptyset} 1 \geq \frac{1}{n! \cdot 2n} \sum_{\pi,k} |P^*(\pi, k)| \geq \frac{n+1}{2n}(\varphi_{G'}(a'_n) + \varphi_{G'}(a''_n)).$$

We conclude that the manipulator cannot gain more than a factor of $2n/(n+1) < 2$ by splitting his weight between two identities.

To see that this bound is tight, consider the game $G = [2, 2, \ldots, 2; 2n]$ and suppose that one of the agents (say, a_n) decides to split into two identities a'_n and a''_n resulting in the game $G' = [2, \ldots, 2, 1, 1; 2n]$. Clearly, in both games the only winning coalition consists of all agents, so we have $\varphi_G(a_n) = 1/n$, $\varphi_{G'}(a'_n) = \varphi_{G'}(a''_n) = 1/(n+1)$, i.e., $\varphi_{G'}(a'_n) + \varphi_{G'}(a''_n) = \frac{2n}{n+1}\varphi_G(a_n)$. □

We have seen that no agent can increase his payoff by more than a factor of 2 by splitting his weight between two identities. In contrast, we will now show that an agent can decrease his payoff by a factor of $\Theta(n)$ by doing so. This shows that a would-be manipulator has to be careful when deciding whether to split his weight, and motivates the algorithmic questions studied in the next two sections.

Theorem 5. *In any weighted voting game, no agent can lower his payoff by more than a factor of $(n+1)/2$ by splitting his weight between two identities. Moreover, there exists a weighted voting game in which splitting into two identities decreases the manipulator's payoff by a factor of $(n+1)/2$.*

Proof. To prove the first part of the theorem, fix a split of a_n into a'_n and a''_n and consider any permutation π of agents in G such that a_n is pivotal for π. It is easy to see that at least one of a'_n and a''_n is pivotal for the permutation $f(\pi)$ obtained from π by replacing a_n with a'_n and a''_n (in this order). Similarly, at least one of a'_n and a''_n is pivotal for the permutation $g(\pi)$ obtained from π by replacing a_n with a''_n and a'_n (in this order). Moreover, all permutations of agents in G' obtained in this manner are distinct, i.e., for any π, π' we have $g(\pi) \neq f(\pi')$, and $\pi \neq \pi'$ implies $f(\pi) \neq f(\pi')$, $g(\pi) \neq g(\pi')$. Consequently, if Π^*_n is the set of all permutations π of the agents in G such that a_n is pivotal for π, and Π^*_{n+1} is the set of all permutations π of the agents in G' such that a'_n or a''_n is pivotal for π, we have $|\Pi^*_{n+1}| \geq 2|\Pi^*_n|$ and

$$\varphi_{G'}(a'_n) + \varphi_{G'}(a''_n) = \frac{|\Pi^*_{n+1}|}{(n+1)!} \geq \frac{2|\Pi^*_n|}{(n+1)!} = \frac{2}{n+1}\varphi_G(a_n).$$

To see that this bound is tight, consider the game $G = [2, 2, \ldots, 2; 2n-1]$ and suppose that one of the agents (say, a_n) decides to split into two identities a'_n and a''_n resulting in

the game $G' = [2, \ldots, 2, 1, 1; 2n-1]$. In the original game G, the only winning coalition consists of all agents, so we have $\varphi_G(a_n) = 1/n$. Now, consider any permutation π of the players in G'. We claim that a'_n is pivotal for π if and only if it appears in the nth position of π, followed by a''_n. Indeed, if $\pi(a'_n) = n$, $\pi(a''_n) = n+1$, then all players in the first $n-1$ positions have weight 2, so $w(S_\pi(a'_n)) = 2n-2$, $w(S_\pi(a'_n) \cup \{a'_n\}) = 2n-1$. Conversely, if $\pi(a'_n) = n+1$, we have $w(S_\pi(a'_n)) = 2n-1 = q$, and if $\pi(a'_n) \leq n-1$, we have $w(S_\pi(a'_n) \cup \{a'_n\}) \leq 2(n-1)$. Finally, if $\pi(a'_n) = n$, but $\pi(a''_n) \neq n+1$, we have $w(S_\pi(a'_n) \cup \{a'_n\}) = 2n-2 < q$. Consequently, a'_n is pivotal for $(n-1)!$ permutations, and, by the same argument, a''_n is also pivotal for (a disjoint set of) $(n-1)!$ permutations. Hence, we have $\varphi_{G'}(a'_n) + \varphi_{G'}(a''_n) = \frac{2(n-1)!}{(n+1)!} = \frac{2}{n+1}\varphi_G(a_n)$. □

5 Complexity of Finding a Beneficial Manipulation

We now examine the problem of finding a beneficial weight split in weighted voting games from the computational perspective. Ideally, the manipulator would like to find a payoff-maximizing split, i.e., a way to split his weight between two or more identities that results in the maximal total payoff. A less ambitious goal is to decide whether there exists a manipulation that increases the manipulator's payoff. However, it turns out that even this problem is computationally hard: in the rest of the section, we will show that checking whether there exists a payoff-increasing split is NP-hard, even if the player is only allowed to use two identities. To formally define the computational problem, we assume that all weights and the quota are integer numbers given in binary.

Definition 6. *An instance of* Beneficial Split *is given by a weighted voting game $G = [w_1, \ldots, w_n; q]$ and a certain target agent a_i. We are asked if there is a way for a_i to split his weight w_i between several new agents $a_i^{(1)}, \ldots, a_i^{(k)}$ so that the sum of their Shapley values is greater than the Shapley value of a_i. Formally, an instance (G, a_i) is a "yes"-instance if there exists a $k \geq 2$ and weights $w_i^{(1)}, \ldots, w_i^{(k)}$ such that $\sum_{j=1,\ldots,k} w_i^{(j)} = w_i$ and in the new game*

$$G' = [w_1, \ldots, w_{i-1}, w_i^{(1)}, \ldots, w_i^{(k)}, w_{i+1}, \ldots, w_n; q]$$

we have $\varphi_{G'}(a_i^{(1)}) + \cdots + \varphi_{G'}(a_i^{(k)}) > \varphi_G(a_i)$.

Remark 7. *Note that we are looking for a* strictly *beneficial manipulation, i.e., one that* increases *the total Shapley value of the manipulator. Indeed, if we were just interested in a split that is not harmful, the problem would always have a trivial solution: by the dummy axiom, assigning a weight of 0 to the new agent does not affect the Shapley value of all agents and therefore is not harmful.*

Our hardness proof is by a reduction from Partition, which is a classical NP-complete problem. An instance of Partition is given by a set of n weights $T = \{t_1, \ldots, t_n\}$. It is a "yes"-instance if it is possible to split T into two subsets $P_1 \subseteq T$, $P_2 \subseteq T$ so that $P_1 \cap P_2 = \emptyset$, $P_1 \cup P_2 = T$, and $\sum_{t_i \in P_1} t_i = \sum_{t_i \in P_2} t_i$, and a "no"-instance if no such partition exists.

The high-level idea of the reduction is as follows. Given an instance of Partition $T = \{t_1, \ldots, t_n\}$, we create a weighted voting game G with $n+2$ agents $a_1, \ldots, a_n, a_x, a_y$, weights $\mathbf{w} = (8t_1, ..., 8t_n, 1, 2)$, and a quota of $q = 4\sum_i t_i + 3$. The weights are chosen so that a_y is a dummy if the original instance of Partition is a "no"-instance, but has some power if it is a "yes"-instance. Moreover, when a partition exists, agent a_y can gain power by splitting into two agents of weight 1 each. It follows that T is a "yes"-instance of Partition if and only if (G, a_y) a "yes"-instance of Beneficial Split. For the rest of the proof, we write $A = \{a_1, \ldots, a_n\}$ and for any $A' \subseteq A$ we set $w(A') = \sum_{i:a_i \in A'} w_i$.

Lemma 8. *If T is a "no"-instance of* Partition*, then agent a_y is a dummy player.*

Proof. Suppose that T is a "no"-instance of Partiton. Consider any $A' \subseteq A$. The set A can be partitioned into two equal-weight subsets if and only if T can, so either $w(A') < w(A)/2$, or $w(A') > w(A)/2$. We will show that in either case a_y cannot be critical to either $A' \cup \{a_y\}$ or $A' \cup \{a_x, a_y\}$, i.e., a_y is dummy player.

The weights of all agents in A are multiples of 8, so $w(A)/2$ is a multiple of 4. Similarly, the weight of A' is a multiple of 8. Hence, if $w(A') < w(A)/2$, it follows that $w(A') \leq w(A)/2 - 4$ and $w(A' \cup \{a_x, a_y\}) \leq w(A)/2 - 4 + 3 < q$. Therefore, $A' \cup \{a_x, a_y\}$ (and *a forteriori* $A' \cup \{a_y\}$) is not a winning coalition, so a_y cannot be pivotal for it.

Similarly, if $w(A') > w(A)/2$, then $w(A') \geq w(A)/2 + 4 > q$, so A' is a winning coalition. Therefore a_y cannot be critical for $A' \cup \{a_y\}$ or $A' \cup \{a_x, a_y\}$. We conclude that by the dummy axiom the Shapley value of a_y is 0. □

Corollary 9. If T is a "no"-instance of Partition, then (G, a_y) is a "no"-instance of Beneficial Split.

Proof. By Lemma 8, if T is a "no"-instance of Partition, the agent a_y is a dummy in G. Now, take any possible split of a_y into several agents $a_y^{(1)}, \ldots, a_y^{(k)}$ and consider any permutation π of the agents in the new game. If there is a $j \leq k$ such that $a_y^{(j)}$ is pivotal for π, then a_y is pivotal for the permutation obtained from π by deleting all agents $a_y^{(l)}$ with $l \neq j$ and replacing $a_y^{(j)}$, with a_y, a contradiction. We conclude that all $a_y^{(j)}$, $j = 1, \ldots, k$ are dummy players and hence their total Shapley value is 0. Therefore, a_y gains no power by splitting and (G, a_y) is a "no"-instance of Beneficial Split. □

Lemma 10. *If T is a "yes"-instance of* Partition*, then a_y can increase his power by splitting into two agents, i.e., (G, a_y) is a "yes"-instance of* Beneficial Split.

Proof. Let T be a "yes"-instance of Partition. Let $\langle P_1, P_2 \rangle$ be a partition of T, so $w(P_1) = w(P_2)$. It corresponds to a partition $\langle A_1, A_2 \rangle$ of A, where $a_i \in A_1$ if and only if $t_i \in P_1$; obviously, we have $w(A_1) = w(A_2)$. We denote $|A_1| = s$, so $|A_2| = n - s$.

It is easy to see that a_y is critical for $A_1 \cup \{a_x, a_y\}$ as well as for $A_2 \cup \{a_x, a_y\}$. There are $(s+1)!(n-s)!$ permutations of $a_1, \ldots, a_n, a_x, a_y$ that put a_y directly after some permutation of $A_1 \cup \{a_x\}$. Similarly, there are $s!(n-s+1)!$ permutations putting a_y directly after some permutation of $A_2 \cup \{a_x\}$. Thus, for each partition $X_i = \langle P_1^i, P_2^i \rangle$, where $|P_1^i| = s$, we have at least $(s+1)!(n-s)! + s!(n-s+1)!$ distinct permutations where a_y is critical. On the other hand, as shown in Lemma 8, if A' is a subset of A such that $w(A') \neq w(A)/2$, then a_y is not critical for $A' \cup \{a_y\}$ or $A' \cup \{a_y, a_x\}$, since either $w(A') \leq w(A)/2 - 4 < q - 3$ or $w(A') \geq w(A)/2 + 4 > q$.

Let $\mathcal{P}$ be the set of all partitions of T, where each partition is counted only once, i.e., $\mathcal{P}$ contains exactly one of the $\langle P_1, P_2 \rangle$ and $\langle P_2, P_1 \rangle$. For each $P_i = \langle P_1^i, P_2^i \rangle \in \mathcal{P}$, we denote $|P_1^i| = s_i$. There is a total of $(n+2)!$ permutations of players in G. Thus, the Shapley value of a_y in G is

$$\varphi_G(a_y) = \sum_{P_i \in \mathcal{P}} \frac{(s_i+1)!(n-s_i)! + s_i!(n-s_i+1)!}{(n+2)!} = \sum_{P_i \in \mathcal{P}} \frac{s_i!(n-s_i)!(s_i+1+n-s_i+1)}{(n+2)!} =$$
$$\sum_{P_i \in \mathcal{P}} \frac{s_i!(n-s_i)!(n+2)}{(n+2)!} = \sum_{P_i \in \mathcal{P}} \frac{s_i!(n-s_i)!}{(n+1)!}.$$

We now consider what happens when a_y splits into two agents, a_y' and a_y'', $w(a_y') = w(a_y') = 1$, resulting in a game $G' = [8t_1, \ldots, 8t_n, 1, 1, 1; \sum_{i=1}^n t_i + 3]$.

Again, let $\langle P_1, P_2 \rangle$, $|P_1| = s_i$, $|P_2| = n - s_i$, be a partition of T, so $w(P_1) = w(P_2)$, and let $\langle A_1, A_2 \rangle$ be the corresponding partition of A. Consider any permutation π which places a_y'' directly after some permutation of $A_1 \cup \{a_x, a_y'\}$; clearly, a_y'' is critical for π. Similarly,

a''_y is critical for any permutation π' which places a''_y directly after some permutation of $A_2 \cup \{a_x, a'_y\}$. There are $(s_i+2)!(n-s_i)!$ permutations putting a''_y directly after some permutation of $A_1 \cup \{a_x, a'_y\}$ and $s_i!(n-s_i+2)!$ permutations putting a''_y directly after some permutation of $A_2 \cup \{a_x, a'_y\}$.

Thus, for each partition $P_i =< P_1^i, P_2^i >$, where $|P_1^i| = s_i$, we have $(s_i+2)!(n-s_i)! + s_i!(n-s_i+2)!$ permutations where a''_y is critical. Switching the roles of a'_y and a''_y, both of which have the same weight and are thus equivalent, we also get that there are $(s_i+2)!(n-s_i)!+s_i!(n-s_i+2)!$ permutations where a'_y is critical. There are $n+3$ agents in G', so there is a total of $(n+3)!$ permutations of the agents. Thus each partition $P_i = (P_1^i, P_2^i)$, $|P_i| = s_i$, contributes $\frac{s_i!(n-s_i)!}{(n+1)!}$ to the Shapley value of a_y in G, and $2\frac{(s_i+2)!(n-s_i)!+s_i!(n-s_i+2)!}{(n+3)!}$ to the sum of the Shapley values of a'_y or a''_y in G'. We will now show that for any partition P_i

$$2\frac{(s_i+2)!(n-s_i)!+s_i!(n-s_i+2)!}{(n+3)!} > \frac{s_i!(n-s_i)!}{(n+1)!}. \tag{2}$$

Summing these inequalities over all partitions P_i will imply $\varphi_{G'}(a'_y) + \varphi_{G'}(a''_y) > \varphi_G(a_y)$, as desired. Inequality (2) can be simplified to

$$2\frac{(s+1)(s+2)+(n-s+1)(n-s+2)}{(n+2)(n+3)} > 1,$$

where we use s instead of s_i to simplify notation, or, equivalently, $2(s+1)(s+2)+2(n-s+1)(n-s+2)-(n+2)(n+3) > 0$. Now, observe that $2(s+1)(s+2)+2(n-s+1)(n-s+2)-(n+2)(n+3) = (n-2s)^2+n+2 > 0$ for any n. This proves inequality (2) for any s, n. It follows that if there is a partition, agent a_y always gains by splitting into two agents of weight 1. Thus, if T is a "yes"-instance of Partition, then (G, a_y) is a "yes"-instance of Beneficial Split. □

We summarize our results in the following theorem.

Theorem 11. Beneficial Split *is NP-hard, even if the only allowed split is into two identities with equal weights.*

Proof. We transform an instance T of Partition into an instance (G, a_y) of Beneficial Split as explained above. We combine Corollary 9 and Lemma 10, and see that T is a "yes" instance of Partition if and only if the (G, a_y) is a "yes"-instance of Beneficial Split. This completes the reduction. □

Remark 12. *Note that we have not shown that* Beneficial Split *is in NP, so we have not proved that it is NP-complete. There are two reasons for this. First, if we allow splits into an arbitrary number of identities, some of the candidate solutions may have exponentially many new agents (e.g., an agent with weight w_i can split into w_i agents of weight 1), or agents whose weights are rational numbers with superpolynomially many digits in their binary representation. Second, even if we circumvent this issue by only considering splits into two identities with integer weights, it is not clear how to verify in polynomial time whether a particular split is beneficial. In fact, since computing the Shapley value in weighted voting games is #P-complete, it is quite possible that* Beneficial Split *is not in NP.*

6 Finding Beneficial Splits

In Section 5, we have shown that it is hard even to test if any beneficial split exists, let alone to find the optimal split. This can be seen as a positive result, since complexity of

finding beneficial splits serves as a barrier for this kind of manipulative behavior. However, it turns out that in many cases manipulators can overcome this difficulty. Indeed, recall that our hardness reduction is from PARTITION. While this problem is NP-hard, its hardness—and hence the hardness of our problem—crucially relies on the fact that the weights of the elements are represented in binary. Indeed, if the weights are given in unary, there is a dynamic programming-based algorithm for PARTITION that runs is time polynomial in size of the input (such algorithms are usually referred to as *pseudopolynomial*). In particular, if all weights are polynomial in n, the running time of this algorithm is polynomial in n. In many natural voting domains the weights of all agents are not too large, so this scenario is quite realistic. It is therefore natural to ask if there exists a pseudopolynomial algorithm for the problem of finding a beneficial split.

It turns out that the answer to this question is indeed positive as long as there is a constant upper bound K on the number of identities that the manipulator can use and all weights are required to be integer. To see this, recall that there is a pseudopolynomial algorithm for computing the Shapley value of any player in a weighted voting game [10]. This algorithm is based on dynamic programming: for any weight W and any $1 \leq k \leq n$, it calculates the number of coalitions of size k that have weight W. One can use the algorithm of [10] to find a beneficial split for a player a_i with weight w_i in a game G as follows. Consider all possible splits $w_i = w_i^{(1)} + \cdots + w_i^{(K)}$, where $w_i^{(j)} \in \mathbb{Z}$, $w_i^{(j)} \geq 0$ for $j = 1, \ldots, K$. Clearly, the number of such splits is at most $(w_i)^K$, which is polynomial in n for constant K. Evaluate the Shapley values of all new agents in any such split and return "yes" if and only if any of these splits results in an increased total payoff. Let $A(G)$ be the running time of the algorithm of paper [10] on instance G. The running time of our algorithm is $O((w_i)^K K \cdot A(G))$, which is clearly pseudopolynomial.

We will now consider a more general setting, where only the weight of the manipulator is polynomially bounded, while the weights of other players can be large. To simplify the presentation, we limit ourselves to the case of two-way splits; however, our approach applies to splits into any constant number of identities. We can use the same high-level approach as in the previous case, i.e., considering all possible splits (because of the weight restriction, there is only polynomially many of them), and computing the Shapley values of both new agents for each split. However, if we were to implement the latter step exactly, it would take exponential time. Therefore, in this version of our algorithm, we replace the algorithm of [10] with an approximation algorithm for computing the Shapley value. Several such algorithms are known: see, e.g., [8, 4]. We will use these algorithms in a black-box fashion. Namely, we assume that we are given a procedure $\texttt{Shapley}(G, a_i, \delta, \epsilon)$ that for any given values of $\epsilon > 0$ and $\delta > 0$ outputs a number v that with probability $1 - \delta$ satisfies $|v - \varphi_G(a_i)| \leq \epsilon$ and runs in time $\text{poly}(n \log w_{\max}, 1/\epsilon, 1/\delta)$. We show how to use this procedure to design an algorithm for finding a beneficial split and relate the performance of our algorithm to that of $\texttt{Shapley}(G, a_i, \delta, \epsilon)$.

Our algorithm is described in Figure 1. It takes a pair of parameters (δ, ϵ) as an input, and uses the procedure $\texttt{Shapley}(G, a_i, \delta, \epsilon)$ as a subroutine. The algorithm outputs "yes" if it finds a split whose total estimated payoff exceeds the payoff of the manipulator in the original game by at least 3ϵ. It can easily be modified to output the optimal split.

Proposition 13. *With probability $1-3\delta$, the output of our algorithm satisfies the following: (i) If the algorithm outputs "yes", then (G, a_i) admits a beneficial integer split; (ii) Conversely, if there is an integer split that increases the payoff to the manipulator by more than 6ϵ, our algorithm outputs "yes". Moreover, the running time of our algorithm is polynomial in nw_i, $1/\epsilon$, and $1/\delta$.*

Proof. Suppose that the algorithm outputs "yes". We have $Prob[v^* < \varphi_G(a_i) - \epsilon] < \delta$, $Prob[v' > \varphi_{G'}(a_i') + \epsilon] < \delta$, $Prob[v'' > \varphi_{G'}(a_i'') + \epsilon] < \delta$. Hence, with probability at least

```
FindSplit(G = [w; q], a_i, δ, ϵ);
v* =Shapley(G, a_i, δ, ϵ);
for j = 0, ..., w_i
    w'_i = j, w''_i = w_i − j;
    G' = [w_1, ..., w_{i−1}, w'_i, w''_i, w_{i+1}, ..., w_n; q];
    v' = Shapley(G', a'_i, δ, ϵ), v'' =Shapley(G', a''_i, δ, ϵ);
    v = v' + v'';
    if v > v* + 3ϵ then return yes;
return no;
```

Figure 1: Algorithm $\texttt{FindSplit}(G = [\mathbf{w}; q], a_i, \delta, \epsilon)$

$1 - 3\delta$, if $v' + v'' > v^* + 3\epsilon$, then $\varphi_{G'}(a'_i) + \varphi_{G'}(a''_i) + 2\epsilon > \varphi_G(a_i) - \epsilon + 3\epsilon$, or, equivalently, $\varphi_{G'}(a'_i) + \varphi_{G'}(a''_i) > \varphi_G(a_i)$.

Conversely, suppose that there is a beneficial split of the form (w'_i, w''_i) that improves player a_i's payoff by at least 6ϵ. As before, with probability at least $1 - 3\delta$ we have that $v^* \leq \varphi_G(a_i) + \epsilon$ and at the step $j = w'_i$ it holds that $v' \geq \varphi_{G'}(a'_i) - \epsilon$, $v'' \geq \varphi_{G'}(a''_i) - \epsilon$. Then $v = v' + v'' \geq \varphi_{G'}(a'_i) + \varphi_{G'}(a''_i) - 2\epsilon > \varphi_G(a_i) + 6\epsilon - 2\epsilon \geq v^* + 3\epsilon$, so the algorithm will output "yes". □

While our algorithm does not *guarantee* finding a successful manipulation, it is possible to control the approximation quality (at the cost of increasing the running time), so that a successful manipulation is found with high probability.

Thus we can see that manipulators have several ways to overcome the computational difficulty of finding the optimal manipulation. Thus, other measures are required to avoid such manipulations.

7 Extensions

In this section, we consider some variants of the model studied in the paper. Our results here are rather preliminary and provide several interesting directions for future research.

Splitting into more than two identities So far, we have mostly discussed the gain (or loss) that an agent can achieve by splitting into two identities. However, it is also possible for an agent to use three or more false names. Potentially, the number of identities an agent can use can be as large as his weight (and if the weights are not required to be integer, it can even be infinite). It would be interesting to see which of our results hold in this more general setting. For example, while our computational hardness result holds for splits into any number of identities, the algorithmic results of the previous section only apply to splits into a constant number of new identities. An obvious open problem here is to design a pseudopolynomial algorithm for finding a beneficial integer split into any number of identities, or to prove that this problem is NP-hard even for small weights (i.e., weights that are polynomial in n). Another question of interest here is to extend the upper and lower bounds of Section 4.1 for this setting.

Manipulation by merging Each situation in which splitting is harmful for an agent directly corresponds to a situation where it is beneficial for several agents to merge, i.e., pretend that they are a single agent whose weight is equal to the total weight of the manipulators.

Some of the results presented in the paper can easily be translated to this domain. In particular, it is not hard to see that the proof of Theorem 11 can be adapted to show that

it is NP-hard to check whether there exists a beneficial merge, and the results of Section 4.1 can be interpreted in terms of merging rather than splitting. However, this problem is very different from the game-theoretic perspective, as it involves coordinated actions by several would-be manipulators who then have to decide how to split the (increased) total payoff. We propose it as a direction for future work.

8 Conclusions

We have considered false-name manipulations in weighted voting games. We have shown that these manipulations can both increase and decrease the manipulator's payoffs, and provided tight upper and lower bounds on the effects of false-name voting. We have also shown that testing whether a beneficial manipulation exists is NP-hard. One may ask why we view this hardness result as an adequate barrier to manipulation, while using Shapley value (which itself is #P-hard to compute) as a payoff division scheme and therefore assuming that it can be computed. To resolve this apparent contradiction, note that the Shapley value corresponds to the voting power, and the players may try to increase their voting power by weight-splitting manipulation even if they cannot compute it. Also, when the Shapley value is used to compute payments, the center, which performs this computation, may have more computational power than individual agents. Furthermore, a payoff division scheme that is based on approximate computation of Shapley value may still be acceptable to the agents, whereas the manipulator may want to know for sure that attempted manipulation will not hurt him (and we have seen that in some cases weight-splitting can considerably decrease the agent's payoffs), or provide him with sufficient benefits to offset the costs of splitting. While the approximation algorithm discussed in the previous section can be used for this purpose, it only works if the manipulator's weight is small. Generalizing it to large weights (i.e., showing that if a beneficial split exists, it can be found by testing a polynomial number of splits) is an interesting open question.

In this paper, we presented results on false-name voting for the case when the payoffs are distributed according to the Shapley value. An obvious research direction is to see if one can derive similar results for other power indices, such as Banzhaf index, as well as other solution concepts used in co-operative games such as, e.g., the nucleolus. More generally, it would be interesting to design a payoff distribution scheme that is resistant to this type of manipulation, or prove that it does not exist.

The study of weighted voting has many applications both in political science and in multiagent systems. There are several possible interpretations for identity-splitting in these contexts, such as obtaining a higher share of the grand coalition's gains when these are distributed according to the Shapley value, or obtaining more political power by splitting a political party into several parties with similar political platforms. In the first case, a false-name manipulation is hard to detect in open anonymous environments, and can thus be very effective. In the second case, the manipulation is done using legitimate tools of political conduct. Therefore, we conjecture that false-name manipulation is widespread in the real world and may become a serious issue in multiagent systems. It is therefore important to develop a better understanding of the effects of this behavior and/or design methods of preventing it.

References

[1] J. F. Banzhaf. Weighted voting doesn't work: a mathematical analysis. *Rutgers Law Review*, 19:317–343, 1965.

[2] V. Conitzer and T. Sandholm. Computing Shapley values, manipulating value division schemes, and checking core membership in multi-issue domains. In *Proc. AAAI'04*, pp. 219–225, 2004.

[3] X. Deng and C. H. Papadimitriou. On the complexity of cooperative solution concepts. *Math. Oper. Res.*, 19(2):257–266, 1994.

[4] S. S. Fatima, M. Wooldridge, and N. R. Jennings. A randomized method for the Shapley value for the voting game. In *Proc. AAMAS'07*, pp. 955–962, 2007.

[5] A. Laruelle. On the choice of a power index. Papers 99-10, Valencia — Instituto de Investigaciones Economicas, 1999.

[6] D. Leech. Voting power in the governance of the international monetary fund. *Annals of Operations Research*, 109(1-4):375–397, 2002.

[7] M. Machover and D. S. Felsenthal. The treaty of Nice and qualified majority voting. *Social Choice and Welfare*, 18(3):431–464, 2001.

[8] I. Mann and L. S. Shapley. Values of large games, IV: Evaluating the electoral college by Monte-Carlo techniques. Technical report, The Rand Corporation, Santa Monica, CA, 1960.

[9] I. Mann and L. S. Shapley. Values of large games, VI: Evaluating the electoral college exactly. Technical report, The Rand Corporation, Santa Monica, CA, 1962.

[10] Y. Matsui and T. Matsui. A survey of algorithms for calculating power indices of weighted majority games. *Journal of the Operations Research Society of Japan*, 43, 2000.

[11] G. Owen. Multilinear extensions and the Banzhaf Value. *Naval Research Logistics Quarterly*, 22(4):741–750, 1975.

[12] L. S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, pp. 31–40, 1953.

[13] L. S. Shapley and M. Shubik. A method for evaluating the distribution of power in a committee system. *American Political Science Review*, 48:787–792, 1954.

[14] P. Straffin. Homogeneity, independence and power indices. *Public Choice*, 30:107–118, 1977.

[15] M. Yokoo, V. Conitzer, T. Sandholm, N. Ohta, A. Iwasaki. Coalitional games in open anonymous environments. In *Proc. AAAI'05*, pp. 509–515, 2005.

[16] M. Yokoo, Y. Sakurai, S. Matsubara. Robust combinatorial auction protocol against false-name bids. *Artificial Intelligence*, 130(2):167.181, 2001.

[17] M. Yokoo, Y. Sakurai, S. Matsubara. The effect of false-name bids in combinatorial auctions: New fraud in Internet auctions. *Games and Economic Behavior*, 46(1):174.188, 2004.

Yoram Bachrach
School of Engineering and Computer Science
Hebrew University
Jerusalem, Israel
Email: `yori@cs.huji.ac.il`

Edith Elkind
School of Electronics and Computer Science
University of Southampton
Southampton, United Kingdom
Email: `ee@ecs.soton.ac.uk`

Computing Kemeny Rankings, Parameterized by the Average KT-Distance

Nadja Betzler, Michael R. Fellows, Jiong Guo, Rolf Niedermeier, and Frances A. Rosamond

Abstract

The computation of Kemeny rankings is central to many applications in the context of rank aggregation. Unfortunately, the problem is NP-hard. Extending our previous work [AAIM 2008], we show that the Kemeny score of an election can be computed efficiently whenever the *average* pairwise distance between two input votes is not too large. In other words, Kemeny Score is fixed-parameter tractable with respect to the parameter "average pairwise Kendall-Tau distance d_a". We describe a fixed-parameter algorithm with running time $O(16^{\lceil d_a \rceil} \cdot \text{poly})$.

1 Introduction

Aggregating inconsistent information has many applications ranging from voting scenarios to meta search engines and fighting spam [1, 4, 5, 7]. In some sense, one deals with *consensus problems* where one wants to find a solution to various "input demands" such that these demands are met as well as possible. Naturally, contradicting demands cannot be fulfilled at the same time. Hence, the consensus solution has to provide a balance between opposing requirements. The concept of *Kemeny consensus* (or Kemeny ranking) is among the most important research topics in this context. In this paper, extending our previous work [3], we study new algorithmic approaches based on parameterized complexity analysis [6, 9, 13] for efficiently computing optimal Kemeny consensus solutions in practically relevant special cases.

Kemeny's voting scheme can be described as follows. An *election* (V, C) consists of a set V of n votes and a set C of m candidates. A vote is a *preference list* of the candidates. For instance, in the case of three candidates a, b, c, the order $c > b > a$ would mean that candidate c is the best-liked and candidate a is the least-liked for this voter. A "Kemeny consensus" is a preference list that is "closest" to the preference lists of the voters. For each pair of votes v, w, the so-called *Kendall-Tau distance* (*KT-distance* for short) between v and w, also known as the number of inversions between two permutations, is defined as

$$\text{KT-dist}(v, w) = \sum_{\{c,d\} \subseteq C} d_{v,w}(c, d),$$

where the sum is taken over all unordered pairs $\{c, d\}$ of candidates, and $d_{v,w}(c, d)$ is 0 if v and w rank c and d in the same order, and 1 otherwise. Using divide-and-conquer, the KT-distance can be computed in $O(m \cdot \log m)$ time [12]. The *score* of a preference list l with respect to an election (V, C) is defined as $\sum_{v \in V} \text{KT-dist}(l, v)$. A preference list l with the minimum score is called a *Kemeny consensus* of (V, C) and its score $\sum_{v \in V} \text{KT-dist}(l, v)$ is the *Kemeny score* of (V, C), denoted as K-score(V, C). The underlying decision problem is as follows:

Kemeny Score
Input: An election (V, C) and a positive integer k.
Question: Is K-score$(V, C) \leq k$?

Known results. We summarize the state of the art concerning the computational complexity of Kemeny Score. Bartholdi et al. [2] showed that Kemeny Score is NP-complete, and it remains so even when restricted to instances with only four votes [7, 8]. Given the computational hardness of Kemeny Score on the one side and its practical relevance on the other side, polynomial-time approximation algorithms have been studied. The Kemeny score can be approximated to a factor of $8/5$ by a deterministic algorithm [15] and to a factor of $11/7$ by a randomized algorithm [1]. Recently, a polynomial-time approximation scheme (PTAS) has been developed [11]. However, the running time is completely impractical and the result is only of theoretical interest. Conitzer, Davenport, and Kalagnanam [5, 4] performed computational studies for the efficient exact computation of a Kemeny consensus, using heuristic approaches such as greedy and branch-and-bound. Hemaspaandra et al. [10] provided further, exact classifications of the classical computational complexity of Kemeny elections. Very recently, we initiated a parameterized complexity study based on various problem parameterizations [3]. We obtained fixed-parameter tractability results for the parameters "score", "number of candidates", "maximum KT-distance between two input votes", and "maximum position range of a candidate".[1] For more details, see Section 2

New results. Our main result is that Kemeny Score can be solved in $O(16^{\lceil d_a \rceil} \cdot \text{poly}(n, m))$ time, where d_a denotes the average KT-distance between the pairs of input votes. This represents a significant improvement of the previous algorithm for the maximum KT-distance between pairs of input votes, which has running time $O((3d_{\max} + 1)! \cdot \text{poly}(n, m))$. Clearly, $d_a \leq d_{\max}$.

1.1 Preliminaries

Let the *position* of a candidate c in a vote v, denoted by $v(c)$, be the number of candidates that are better than c in v. That is, the leftmost (and best) candidate in v has position 0 and the rightmost has position $m - 1$. For an election (V, C) and a candidate $c \in C$, the *average position* $p_a(c)$ of c is defined as

$$p_a(c) := \frac{1}{n} \cdot \sum_{v \in V} v(c).$$

For an election (V, C) the average KT-distance d_a is defined as

$$d_a := \frac{1}{n(n-1)} \cdot \sum_{\{u,v\} \in V, u \neq v} \text{KT-dist}(u, v).$$

Note that an equivalent definition is given by

$$d_a := \frac{1}{n(n-1)} \cdot \sum_{a,b \in C} \#v(a > b) \cdot \#v(b > a),$$

where for two candidates a and b the number of input votes in which a is ranked better than b is denoted by $\#v(a > b)$. This definition is useful if the input is provided by the outcomes of the pairwise elections of the candidates including the margins of victory.

We briefly introduce the relevant notions of parameterized complexity theory [6, 9, 13]. Parameterized algorithmics aims at a multivariate complexity analysis of problems. This is done by studying relevant problem parameters and their influence on the computational complexity of problems. The hope lies in accepting the seemingly inevitable combinatorial

[1] The parameterization by position range has only been discussed in the long version of [3].

explosion for NP-hard problems, but confining it to the parameter. Hence, the decisive question is whether a given parameterized problem is *fixed-parameter tractable (FPT)* with respect to the parameter, often denoted k. In other words, for an input instance I together with the parameter k, we ask for the existence of a solving algorithm with running time $f(k) \cdot \text{poly}(|I|)$ for some computable function f.

2 Parameterizations of the Kemeny Score Problem

In recent work [3], we initiated a parameterized complexity study of Kemeny Score. In this section, we review the considered parameterizations and results.

An election can be interpreted as having (at least) two "dimensions", the set of votes and the set of candidates. Thus, the "number n of votes" and the "number m of candidates" lead to natural parameterizations. Fixed-parameter tractability with respect to the parameter "number of votes" would imply P=NP since Kemeny Score is already NP-complete for instances with four votes [7]. In contrast, concerning m as a parameter, there is a trivial algorithm that tests all $m!$ orderings of the candidates for a Kemeny consensus. Using a dynamic programming approach, we were able to lower the combinatorial explosion in m; more specifically, we provided an exact algorithm running in $O(2^m \cdot m^2 n)$ time [3].[2]

A common parameterization in parameterized algorithmics is the size of the solution of a problem, motivating the consideration of "Kemeny score k" as a parameter. Using the Kemeny score as a parameter, a preprocessing procedure (that is, a "problem kernelization") together with a search tree approach led to a fixed-parameter algorithm that runs in $O(1.53^k + m^2 n)$ time. The drawback of this parameterization is that the Kemeny score can become large for many instances.

Finally, we turned our attention to two structural parameterizations: the "maximum range of candidate positions" and the "maximum KT-distance".[3] For an election (V, C), the maximum range r of candidate positions is defined as

$$r := \max_{v,w \in V, c \in C} \{|v(c) - w(c)|\}.$$

For this parameterization, we developed a dynamic programming algorithm that is based on the observation that we can "decompose" the input votes into two parts. The resulting running time is $O((3r+1)! \cdot r \log r \cdot mn)$. Further, the maximum KT-distance $d_{\max}$ is defined as

$$d_{\max} := \max_{v,w \in V} \text{KT-dist}(v, w).$$

We showed that for every election we have $r \leq d_{\max}$. Thus, our results for the maximum range of candidate positions also hold for the maximum distance. The parameterization by $d_{\max}$ was the main reason to study the parameterization by the maximum range of candidate positions and, further, motivated us to consider the average distance as parameter in this work.

In our previous work [3], we also extended some of our findings to two generalizations of Kemeny Score; in one case allowing ties and in the other case dealing with incomplete information. For more details, we refer to there [3].

[2] In a different context, this result has been independently achieved by Raman et al. [14].

[3] In the conference version of [3] we only dealt with the maximum KT-distance as parameter whereas in the full version (invited for submission to a special issue of *Theoretical Computer Science*) we discussed both structural parameterizations as we do here.

3 Parameter "Average KT-Distance"

In this section, we further extend the range of parameterizations studied so far by giving a fixed-parameter algorithm with respect to the parameter "average KT-distance". We start with showing how the average KT-distance can be used to upper-bound the range of positions that a candidate can take in any optimal Kemeny consensus. Based on this crucial observation, we then state the algorithm.

3.1 A Crucial Observation

Our fixed-parameter tractability result with respect to the average KT-distance of the input is based on the following lemma.

Lemma 1. *Let d_a be the average KT-distance of an election (V, C) and $d := \lceil d_a \rceil$. Then, in every optimal Kemeny consensus l, for every candidate $c \in C$ with respect to its average position $p_a(c)$ we have $p_a(c) - d < l(c) < p_a(c) + d$.*

Proof. The proof is by contradiction and consists of two claims: First, we show that we can find a vote with Kemeny score less than $d \cdot n$, that is, the Kemeny score of the instance is upper-bounded by $d \cdot n$. Second, we show that in every Kemeny consensus every candidate is in the claimed range. More specifically, we prove that every consensus in which the position of a candidate is not in a "range d of its average position" has a Kemeny score greater than $d \cdot n$, a contradiction to the first claim.

Claim 1: K-score$(V, C) < d \cdot n$.

Proof of Claim 1: To prove Claim 1, we show that there is a vote $v \in V$ with $\sum_{w \in V} \text{KT-dist}(v, w) < d \cdot n$, implying this upper bound for an optimal Kemeny consensus as well. By definition,

$$d_a = \frac{1}{n(n-1)} \cdot \sum_{\{v,w\} \in V, v \neq w} \text{KT-dist}(v, w) \tag{1}$$

$$\Rightarrow \exists v \in V \text{ with } d_a \geq \frac{1}{n(n-1)} \cdot n \cdot \sum_{w \in V} \text{KT-dist}(v, w) = \frac{1}{n-1} \cdot \sum_{w \in V} \text{KT-dist}(v, w) \tag{2}$$

$$\Rightarrow \exists v \in V \text{ with } d_a \cdot n > \sum_{w \in V} \text{KT-dist}(v, w). \tag{3}$$

Since we have $d = \lceil d_a \rceil$, Claim 1 follows directly from Inequality (3).

The next claim shows the given bound on the range of possible candidates positions.

Claim 2: In every optimal Kemeny consensus l, every candidate $c \in C$ fulfills $p_a(c) - d < l(c) < p_a(c) + d$.

Proof of Claim 2: We start by showing that, for every candidate $c \in C$ we have

$$\text{K-score}(V, C) \geq \sum_{v \in V} |l(c) - v(c)|. \tag{4}$$

Note that, for every candidate $c \in C$, for two votes v, w we must have KT-dist$(v, w) \geq |v(c) - w(c)|$. Without loss of generality, assume that $v(c) > w(c)$. Then, there must be at least $v(c) - w(c)$ candidates that have a smaller position than c in v and that have a greater position than c in w. Further, each of these candidates increases the value of KT-dist(v, w)

by one. Based on this, Inequality (4) directly follows as, by definition, K-score$(V, C) = \sum_{v \in V} \text{KT-dist}(v, l)$.

To simplify the proof of Claim 2, in the following, we shift the positions in l such that $l(c) = 0$. Accordingly, we shift the positions in all votes in V, that is, for every $v \in V$ and every $a \in C$, we decrease $v(a)$ by the original value of $l(c)$. Clearly, shifting all positions does not affect the relative differences of positions between two candidates. Then, let the set of votes in which c has a nonnegative position be V^+ and let V^- denote the remaining set of votes, that is, $V^- := V \backslash V^+$.

Now, we show that if candidate c is placed outside of the given range in an optimal Kemeny consensus l, then K-score$(V, C) > d \cdot n$. The proof is by contradiction. We distinguish two cases:

Case 1: $l(c) \geq p_a(c) + d$.
As $l(c) = 0$, in this case $p_a(c)$ becomes negative. Then,

$$0 \geq p_a(c) + d \Leftrightarrow -p_a(c) \geq d.$$

It follows that $|p_a(c)| \geq d$. The following shows that Claim 2 holds for this case.

$$\sum_{v \in V} |l(c) - v(c)| = \sum_{v \in V} |v(c)| = \sum_{v \in V^+} |v(c)| + \sum_{v \in V^-} |v(c)|. \tag{5}$$

Next, replace the term $\sum_{v \in V^-} |v(c)|$ in (5) by an equivalent term that depends on $|p_a(c)|$ and $\sum_{v \in V^+} |v(c)|$. For this, use the following, derived from the definition of $p_a(c)$:

$$\begin{aligned} n \cdot p_a(c) &= \sum_{v \in V^+} |v(c)| - \sum_{v \in V^-} |v(c)| \\ \Leftrightarrow \sum_{v \in V^-} |v(c)| &= n \cdot (-p_a(c)) + \sum_{v \in V^+} |v(c)| = n \cdot |p_a(c)| + \sum_{v \in V^+} |v(c)|. \end{aligned}$$

The replacement results in

$$\sum_{v \in V} |l(c) - v(c)| = 2 \cdot \sum_{v \in V^+} |v(c)| + n \cdot |p_a(c)| \geq n \cdot |p_a(c)| \geq n \cdot d.$$

This says that K-score$(V, C) \geq n \cdot d$, a contradiction to Claim 1.

Case 2: $l(c) \leq p_a(c) - d$.
Since $l(c) = 0$, the condition is equivalent to $0 \leq p_a(c) - d \Leftrightarrow d \leq p_a(c)$, and we have that $p_a(c)$ is nonnegative. Now, we show that Claim 2 also holds for this case.

$$\begin{aligned} \sum_{v \in V} |l(c) - v(c)| = \sum_{v \in V} |v(c)| &= \sum_{v \in V^+} |v(c)| + \sum_{v \in V^-} |v(c)| \\ &\geq \sum_{v \in V^+} v(c) + \sum_{v \in V^-} v(c) = p_a(c) \cdot n \geq d \cdot n. \end{aligned}$$

Thus, also in this case K-score$(V, C) \geq n \cdot d$, a contradiction to Claim 1. □

Based on Lemma 1, for every position we can define the set of candidates that can take this position in an optimal Kemeny consensus. The subsequent definition will be useful for the formulation of the algorithm.

Definition 1. Let (V, C) be an election. For $i \in \{0, \ldots, m-1\}$, let P_i denote the set of candidates that can assume the position i in an optimal Kemeny consensus, that is, $P_i := \{c \in C \mid p_a(c) - d < i < p_a(c) + d\}$.

Based on Lemma 1, we can easily show the following.

Lemma 2. *For every position i, the size of P_i is at most $4d$.*

Proof. The proof is by contradiction. Assume that there is a position i with $|P_i| > 4d$. Due to Lemma 1, for every candidate $c \in P_i$ the positions which c may assume in an optimal Kemeny consensus can differ by at most $2d-1$. This is true because, otherwise, candidate c could not be in the given range around its average position. Then, in a Kemeny consensus, each of the at least $4d+1$ candidates must hold a position that differs at most by $2d-1$ from position i. As there are only $4d-1$ such positions ($2d-1$ on the left and $2d-1$ on the right of i), one obtains a contradiction. □

3.2 Basic Idea of the Algorithm

In Subsection 3.4, we will present a dynamic programming algorithm for Kemeny Score. It exploits the fact that every candidate can only appear in a fixed range of positions in an optimal Kemeny consensus.[4] The algorithm "generates" a Kemeny consensus from the left to the right. It tries out all possibilities for ordering the candidates locally and then combines these local solutions to yield a Kemeny consensus.

More specifically, according to Lemma 2 the number of candidates that can take a position i in an optimal Kemeny consensus for any $0 \leq i \leq m-1$ is at most $4d$. Thus, for position i, we can test all possible candidates. Having chosen a candidate for position i, the remaining candidates that could also assume i must either be left or right of i in a Kemeny consensus. Thus, we test all possible two-partitionings of this subset of candidates and compute a "partial" Kemeny score for every possibility. For the computation of the partial Kemeny scores at position i we make use of the partial solutions computed for the previous position $i-1$.

3.3 Definitions for the Algorithm

To state the algorithm, we need some further definitions. For $i \in \{0, \ldots, m-1\}$, let $I(i)$ denote the set of candidates that could be "inserted" at position i *for the first time*, that is,

$$I(i) := \{c \in C \mid c \in P_i \text{ and } c \notin P_{i-1}\}.$$

Let $F(i)$ denote the set of candidates that must be "forgotten" at latest at position i, that is,

$$F(i) := \{c \in C \mid c \notin P_i \text{ and } c \in P_{i-1}\}.$$

For our algorithm, it is essential to subdivide the overall Kemeny score into *partial Kemeny scores* (pK). More precisely, for a candidate c and a subset of candidates R with $c \notin R$, we set

$$\mathrm{pK}(c, R) := \sum_{c' \in R} \sum_{v \in V} d_v^R(c, c'),$$

where for $c \notin R$ and $c' \in R$ we have $d_v^R(c, c') := 0$ if in v we have $c' > c$, and $d_v^R(c, c') := 1$, otherwise. Intuitively, the partial Kemeny score denotes the score that is "induced" by

[4] In contrast, the previous dynamic programming algorithms from [3] for the parameters "maximum range of candidate positions" and "maximum KT-distance" rely on decomposing the input. Further, here we obtain a much better running time by using a more involved dynamic programming approach.

candidate c and the candidate subset R if the candidates of R have greater positions than c in an optimal Kemeny consensus.[5] Then, for a Kemeny consensus $l := c_0 > c_1 > \cdots > c_{m-1}$, the overall Kemeny score can be expressed by partial Kemeny scores as follows.

$$\text{K-score}(V, C) = \sum_{i=0}^{m-2} \sum_{j=i+1}^{m-1} \sum_{v \in V} d_{v,l}(c_i, c_j) \tag{6}$$

$$= \sum_{i=0}^{m-2} \sum_{c' \in R} \sum_{v \in V} d_v^R(c_i, c') \text{ for } R := \{c_j \mid i < j < m\} \tag{7}$$

$$= \sum_{i=0}^{m-2} \text{pK}(c_i, \{c_j \mid i < j < m\}). \tag{8}$$

Next, consider the three-dimensional dynamic programming table. Roughly speaking, define an entry for every position i, every candidate c that can assume i, and every candidate subset C' of $P_i \backslash \{c\}$. The entry stores the "minimum partial Kemeny score" over all possible orders of the candidates of C' under the condition that c takes position i and all candidates of C' take positions smaller than i. To define the dynamic programming table formally, we need some further notation.

Let $\Pi(C')$ denote the set of all possible orders of the candidates in C', where $C' \subseteq C$. Further, consider a Kemeny consensus in which every candidate of C' has a position smaller than every candidate in $C \backslash C'$. Then, the *minimum partial Kemeny score restricted to* C' is defined as

$$\min_{(c_1 > c_2 > \cdots > c_x) \in \Pi(C')} \left\{ \sum_{s=1}^{x} \text{pK}(c_s, \{c_j \mid s < j < m\} \cup (C \backslash C')) \right\} \text{ with } x := |C'|.$$

That is, it denotes the minimum partial Kemeny score over all orders of C'. We define an entry of the dynamic programming table T for a position i, a candidate $c \in P_i$, and a candidate subset $P_i' \subseteq P_i$ with $c \notin P_i'$. For this, we define $L := \bigcup_{j \leq i} F(j) \cup P_i'$. Then, an entry $T(i, c, P_i')$ denotes the minimum partial Kemeny score restricted to the candidates in $L \cup \{c\}$ under the assumptions that c is at position i in a Kemeny consensus, all candidates of L have positions smaller than i, and all other candidates have positions greater than i. That is, for $|L| = i - 1$, define

$$T(i, c, P_i') := \min_{(c_0 > \cdots > c_{i-1}) \in \Pi(L)} \sum_{s=0}^{i-1} \text{pK}(c_s, C \backslash \{c_j \mid j \leq s\}) + \text{pK}(c, C \backslash (L \cup \{c\})).$$

3.4 Dynamic Programming Algorithm

The algorithm is displayed in Fig. 1. It is easy to modify the algorithm such that it outputs an optimal Kemeny consensus: for every entry $T(i, c, P_i')$, one additionally has to store a candidate c' that minimizes $T(i-1, c', (P_i' \cup F(i)) \backslash \{c'\})$ in line *11*. Then, starting with a minimum entry for position $m - 1$, we reconstruct a Kemeny consensus by iteratively adding the "predecessor" candidate. The asymptotic running time remains unchanged. Moreover, in several applications, it is helpful not having *one* optimal Kemeny consensus but to enumerate all of them. At the expense of an increased running time, our algorithm can be extended to provide such an enumeration by storing all possible predecessor candidates.

Lemma 3. *The algorithm in Fig. 1 correctly computes* Kemeny Score.

[5] By convention and somewhat counterintuitive, we say that candidate c has a greater position than candidate c' if $c' > c$ in a vote.

Input: An election (V, C) and, for every $0 \leq i < m$, the set P_i of candidates that can assume position i in an optimal Kemeny consensus.
Output: The Kemeny score of (V, C).

Initialization:
01 **for** $i = 0, \ldots, m-1$
02 **for all** $c \in P_i$
03 **for all** $P'_i \subseteq P_i \backslash \{c\}$
04 $T(i, c, P'_i) := +\infty$
05 **for all** $c \in P_0$
06 $T(0, c, \emptyset) := \mathrm{pK}(c, C \backslash \{c\})$

Update:
07 **for** $i = 1, \ldots, m-1$
08 **for all** $c \in P_i$
09 **for all** $P'_i \subseteq P_i \backslash \{c\}$
10 **if** $|P'_i \cup \bigcup_{j \leq i} F(j)| = i - 1$ and $T(i-1, c', (P'_i \cup F(i)) \backslash \{c'\})$ is defined **then**

11
$$T(i, c, P'_i) = \min_{c' \in P'_i \cup F(i)} T(i-1, c', (P'_i \cup F(i)) \backslash \{c'\}) + \mathrm{pK}(c, (P_i \cup \bigcup_{i<j<m} I(j)) \backslash (P'_i \cup \{c\}))$$

Output:
12 K-score $= \min_{c \in P_{m-1}} T(m-1, c, P_{m-1} \backslash \{c\})$

Figure 1: Dynamic programming algorithm for Kemeny Score

Proof. For the correctness, we have to show two points:

First, all table entries are well-defined, that is, for an entry $T(i, c, P'_i)$ concerning position i there must be exactly $i-1$ candidates that have positions smaller than i. This condition is assured by line *10* of the algorithm.[6]

Second, we must ensure to find an optimal solution. Due to Equality (8), we know that the Kemeny score can be decomposed into partial Kemeny scores. Thus, it remains to show that the algorithm considers a decomposition that leads to an optimal solution. For every position the algorithm tries all candidates in P_i. According to Lemma 1, one of these candidates must be the "correct" candidate c for this position. Further, for c we can show that the algorithm tries a sufficient set of possibilities to partition all remaining candidates $C\backslash\{c\}$ such that they have either smaller or greater positions than i. More precisely, every candidate of $C\backslash\{c\}$ must be in exactly one of the following three subsets:

1. The set F of candidates that have already been forgotten, that is, $F := \bigcup_{0\leq j\leq i} F(j)$,
2. the set of candidates that can assume position i, that is, $P_i\backslash\{c\}$, or
3. the set I of candidates that are not inserted yet, that is, $I := \bigcup_{i<j<m} I(j)$.

Due to Lemma 1 and the definition of $F(j)$, we know that a candidate of F cannot take a position greater than $i-1$ in an optimal Kemeny consensus. Thus, it is sufficient to try only partitions in which the candidates of F have positions smaller than i. Analogously, one can argue that for all candidates in I it is sufficient to consider partitions in which they have positions greater than i. Thus, it remains to try all possibilities to partition the candidates of P_i. This is done in line *09* of the algorithm. Thus, the algorithm returns an optimal Kemeny score. □

Theorem 1. Kemeny Score *can be solved in* $O(n^2 \cdot m \log m + 16^d \cdot (16d^2 \cdot m + 4d \cdot m^2 \log m \cdot n))$ *time with average KT-distance* d_a *and* $d := \lceil d_a \rceil$. *The size of the dynamic programming table is* $O(16^d \cdot 4dm)$.

Proof. The dynamic programming procedure requires the set of candidates P_i for $0 \leq i < m$ as input. To determine P_i for all $0 \leq i < m$, we need the average positions of all candidates and the average KT-distance d_a of (V, C). To determine d_a, we compute the pairwise distances of all pairs of votes. As we have $O(n^2)$ pairs and the pairwise KT-distance can be computed in $O(m \log m)$ time [12], this takes $O(n^2 \cdot m \log m)$ time. The average positions of all candidates can be computed in $O(n \cdot m)$ time by iterating once over every vote and adding the position of every candidate to a counter variable for this candidate. Thus, the input for the dynamic programming algorithm can be provided in $O(n^2 \cdot m \log m)$ time.

Concerning the dynamic programming algorithm itself, due to Lemma 2, for $0 \leq i < m$, the size of P_i is upper-bounded by $4d$. Then, for the initialization as well as for the update, the algorithm iterates over m positions, $4d$ candidates, and 2^{4d} candidates subsets. Whereas the initialization in the innermost step (line *04*) can be done in constant time, in every innermost step of the update phase (line *11*) we have to look for a minimum entry and we have to compute a pK-score. To find the minimum, we have to consider all candidates of $P'_i \cup F(i)$. As $P'_i \cup F(i)$ is a subset of P_{i-1}, it can contain at most $4d$ candidates. Further, the required pK-score can be computed in $O(n \cdot m \log m)$ time. Thus, for the dynamic programming we arrive at the running time of $O(m \cdot 4d \cdot 2^{4d} \cdot (4d + n \cdot m \log m)) = O(16^d \cdot (16d^2 \cdot m + 4d \cdot m^2 \log m \cdot n))$.

[6] It can still happen that a candidate takes a position outside of the required range around its average position. Since such an entry cannot lead to an optimal solution according to Lemma 1, this does not affect the correctness of the algorithm. To improve the running time it would be convenient to "cut away" such possibilities. We defer considerations in this direction to an extended version of this paper.

Concerning the size of the dynamic programming table, there are m positions and at most $4d$ candidates that can assume a position. The number of considered subsets is bounded from above by 2^{4d}. Hence, the size of T is $O(16^d \cdot 4d \cdot m)$. □

Finally, let us discuss the differences between the dynamic programming algorithm we used for the "maximum range of candidate positions" in [3] and the algorithm presented in this work. In our previous work [3], the dynamic programming table stored all possible orders of the candidates of a given subset of candidates. In this work, we eliminate the need to store all orders by using the decomposition of the Kemeny score into partial Kemeny scores. This allows us to restrict the considerations for a position to a candidate and its order relative to all other candidates. We believe that our new approach can also be used to improve the running time of the algorithm of [3].

4 Conclusion

We significantly improved the running time for a natural parameterization (maximum KT-distance between two input votes) for the Kemeny Score problem. There have been some experimental studies [5, 4] that hinted that the Kemeny problem is easier when the votes are close to a consensus (and thus tend to have a small average distance). Our results for the average distance parameterization can be regarded as a theoretical explanation for this behavior.

As further challenges for future work, we envisage the following:

- Extend our findings to the Kemeny Score problem with input votes that may have ties or that may be incomplete (also see [3]).
- Extend our results to improve the running time for the parameterization by position range—we conjecture that this is not hard to do.
- Improve the running time as well as the memory consumption (which is exponential in the parameter)—we believe that significant improvements are still possible.
- Implement the algorithms for the parameters "number of candidates", "range of position of candidates" [3], and "average KT-distance" (including some maybe heuristic improvements of the running times).
- Investigate typical values of the average KT-distance, either under some distributional assumption or for real-world data.

Acknowledgements. We are grateful to an anonymous referee of *COMSOC 2008* for constructive feedback. This work was supported by the DFG, research project DARE, GU 1023/1, Emmy Noether research group PIAF, NI 369/4, and project PALG, NI 369/8 (Nadja Betzler and Jiong Guo). Michael R. Fellows and Frances A. Rosamond were supported by the Australian Research Council. This work was done while Michael Fellows stayed in Jena as a recipient of the Humboldt Research Award of the Alexander von Humboldt foundation, Bonn, Germany.

Nadja Betzler, Jiong Guo, and Rolf Niedermeier,
Institut für Informatik,
Friedrich-Schiller-Universität Jena,
Ernst-Abbe-Platz 2,
D-07743 Jena, Germany.
Email: `(betzler,guo,niedermr)@minet.uni-jena.de`

Michael R. Fellows and Frances A. Rosamond,
PC Research Unit, Office of DVC (Research),
University of Newcastle,
Callaghan, NSW 2308, Australia.
Email: (michael.fellows,frances.rosamond)@newcastle.edu.au

References

[1] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: Ranking and clustering. In *Proc. 37th STOC*, pages 684–693. ACM, 2005.

[2] J. Bartholdi III, C. A. Tovey, and M. A. Trick. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare*, 6:157–165, 1989.

[3] N. Betzler, M. R. Fellows, J. Guo, R. Niedermeier, and F. A. Rosamond. Fixed-parameter algorithms for Kemeny scores. In *Proc. of 4th AAIM*, volume 5034 of *LNCS*, pages 60–71. Springer, 2008. Long version submitted to *Theoretical Computer Science*.

[4] V. Conitzer, A. Davenport, and J. Kalagnanam. Improved bounds for computing Kemeny rankings. In *Proc. 21st AAAI*, pages 620–626, 2006.

[5] A. Davenport and J. Kalagnanam. A computational study of the Kemeny rule for preference aggregation. In *Proc. 19th AAAI*, pages 697–702, 2004.

[6] R. G. Downey and M. R. Fellows. *Parameterized Complexity*. Springer, 1999.

[7] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the Web. In *Proc. of 10th WWW*, pages 613–622, 2001.

[8] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation revisited, 2001. Manuscript.

[9] J. Flum and M. Grohe. *Parameterized Complexity Theory*. Springer, 2006.

[10] E. Hemaspaandra, H. Spakowski, and J. Vogel. The complexity of Kemeny elections. *Theoretical Computer Science*, 349:382–391, 2005.

[11] C. Kenyon-Mathieu and W. Schudy. How to rank with few errors. In *Proc. 39th STOC*, pages 95–103. ACM, 2007.

[12] J. Kleinberg and E. Tardos. *Algorithm Design*. Addison Wesley, 2006.

[13] R. Niedermeier. *Invitation to Fixed-Parameter Algorithms*. Oxford University Press, 2006.

[14] V. Raman and S. Saurabh. Improved fixed parameter tractable algorithms for two "edge" problems: MAXCUT and MAXDAG. *Information Processing Letters*, 104(2):65–72, 2007.

[15] A. van Zuylen and D. P. Williamson. Deterministic algorithms for rank aggregation and other ranking and clustering problems. In *Proc. 5th WAOA*, volume 4927 of *LNCS*, pages 260–273. Springer, 2007.

Three-sided stable matchings with cyclic preferences and the kidney exchange problem[1]

Péter Biró and Eric McDermid

Abstract

Knuth [14] asked whether the stable matching problem can be generalised to three dimensions i. e., for families containing a man, a woman and a dog. Subsequently, several authors considered the three-sided stable matching problem with cyclic preferences, where men care only about women, women only about dogs, and dogs only about men. In this paper we prove that if the preference lists may be incomplete, then the problem of deciding whether a stable matching exists, given an instance of three-sided stable matching problem with cyclic preferences is NP-complete. Considering an alternative stability criterion, strong stability, we show that the problem is NP-complete even for complete lists. These problems can be regarded as special types of stable exchange problems, therefore these results have relevance in some real applications, such as kidney exchange programs.

1 Introduction

An instance of the Stable Marriage problem (SM) comprises a set of n men $a_1, \ldots, a_n$ and a set of n women $b_1, \ldots, b_n$. Each person has a complete preference list consisting of the members of the opposite sex. If b_j precedes b_k on a_i's list then a_i is said to *prefer* b_j to b_k. The problem is to find a matching that is *stable* in the sense that no man and woman both prefer each other to their current partner in the matching. The Stable Marriage problem was introduced by Gale and Shapley [9]. They constructed a linear time algorithm that always finds a stable matching for an SM instance.

Considering the Stable Marriage problem with Incomplete Lists (SMI), the only difference is that the numbers of men and women are not necessarily equal and each preference list consist of a subset of the members of the opposite sex, i.e., each person lists his or her *acceptable partners*. Here, a matching $\mathcal{M}$ is a set of acceptable pairs, and $\mathcal{M}$ is stable if for every pair $(a_i, b_j) \notin \mathcal{M}$, either a_i prefers his matching partner $\mathcal{M}(a_i)$ to b_j or b_j prefers her matching partner $\mathcal{M}(b_j)$ to a_i. We can model this problem by a bipartite graph $G = (A \cup B, E)$, where the sets of vertices, A and B, correspond to the sets of men and women, respectively, and the set of edges, E represents the acceptable pairs. An extended version of the Gale–Shapley algorithm always produces a stable matching for this setting too.

In an instance of the Stable Marriage problem with Ties and Incomplete Lists (SMTI) it is possible that an agent is indifferent between some acceptable agents from the opposite set; in such a case, these agents appear together in a *tie* in the preference list. Here, a matching $\mathcal{M}$ is stable if there is no blocking pair $(a_i, b_j) \notin \mathcal{M}$ such that a_i is either unmatched or prefers b_j to $\mathcal{M}(a_i)$, and simultaneously b_j is either unmatched or prefers a_i to $\mathcal{M}(b_j)$. Manlove et al. [15] proved that the problem of finding a stable matching of maximum cardinality for an instance of SMTI, the so-called MAX SMTI problem, is NP-hard.

The Three-Dimensional Stable Matching problem (3DSM), also referred to as the Three Gender Stable Marriage problem, was introduced by Knuth [14]. Here, we have three sets of agents: men, women and dogs, say, and each agent has preference over all pairs from

[1]This work was supported by EPSRC grant EP/E011993/1.The first author was supported also by OTKA grant K69027.

the two other sets. A *matching* is a set of disjoint *families* i.e., triples of the form (man, woman, dog). A matching is *stable* if there exists no blocking family that is preferred by all its members to their current families in the matching.

Alkan [2] gave the first example of an instance of 3DSM where no stable matching exists. Ng and Hirschberg [17] proved that the problem of deciding whether a stable matching exists, given an instance of 3DSM, is NP-complete; later Subramanian [26] gave an alternative proof for this. Recently, Huang [10] proved that the problem remains NP-complete even if the preference lists are "consistent". (A preference list is inconsistent if, for example, man m ranks (w_1, d_1) higher than (w_2, d_1), but he also ranks (w_2, d_2) higher than (w_1, d_2), so he does not consistently prefer woman w_1 to woman w_2.)

As an open problem, Ng and Hirschberg [17] mentioned the cyclic 3DSM, defined formally in Section 2, where men only care about women, women only care about dogs and dogs only care about men. Boros et al. [5] showed that if the number of agents n, is at most 3 in every set, then a stable matching always exists. Eriksson et al. [8] proved that this also holds for $n = 4$ and conjectured that a stable matching exists for every instance of cyclic 3DSM.

In Section 2, we study the cyclic 3DSM problem with Incomplete Lists (cyclic 3DSMI). Here, each preference list may consist of a subset of the members of the next gender, i.e. his, her or its *acceptable partners*, and the cardinalities of the sets are not necessarily the same, a matching is a set of acceptable families. Thus cyclic 3DSMI is obtained via a natural generalisation of cyclic 3DSM in a way analogous to the extension SMI of SM. First we give an instance of cyclic 3DSMI for $n = 6$ where no stable matching exists. Then, by using this instance as a gadget, we show that the problem of deciding whether a stable matching exists in an instance of cyclic 3DSMI is NP-complete. We reduce from MAX SMTI.

In Section 3, we study the cyclic 3DSM problem under *strong stability*. A matching is strongly stable if there exists no *weakly blocking family*. This is a family not in the matching that is weakly preferred by all its members (i.e. no member prefers his original family to the new blocking family). We show that the problem of deciding whether a strongly stable matching exists in an instance of cyclic 3DSM is NP-complete.

In Section 4, we describe the correspondence between the cyclic 3DSMI problem and the so-called stable exchange problem with restrictions, defined in Section 4. More precisely, we show that the 3-way stable 3-way exchange problem for tripartite cyclic graphs is equivalent to cyclic 3DSMI. Therefore, the complexity result for cyclic 3DSMI applies also to the 3-way stable 3-way exchange problem, which is an important model for the kidney exchange problem (this application is described in further detail in Section 4).

We remark that all of these problems (namely, SM, SMI, 3DSM, 3DSMI, cyclic 3DSM and cyclic 3DSMI) can be considered as special *coalition formation games*, where the notion of a stable matching is equivalent to the notion of a *core element* in the corresponding NTU-game. Those games, where the set of *basic coalitions* contain all singletons (i.e. where every player has the right not to cooperate with the others) correspond to the stable matching problems with incomplete lists. See more about this correspondence in [3].

2 Cyclic 3DSMI is NP-complete

Problem definition

We consider three sets of agents: M, W, D (men, women and dogs). Every man has a strict preference list over the women that are acceptable to him. Analogously, every woman has a strict preference list over her acceptable dogs, and every dog has a strict preference list over its acceptable men. The list of an agent x is denoted by $P(x)$. A *matching* $\mathcal{F}$ is a set of disjoint families, i.e., triples from $M \times W \times D$, such that for each family $(m, w, d) \in \mathcal{F}$, w is acceptable to m, d is acceptable to w and m is acceptable to d. Formally, if $(m, w, d) \in \mathcal{F}$,

then we say that $\mathcal{F}(m) = w$, $\mathcal{F}(w) = d$ and $\mathcal{F}(d) = m$, thus in a matching, $\mathcal{F}(x) \in P(x) \cup \{x\}$ holds for every agent x, where $\mathcal{F}(x) = x$ means that agent x is unmatched in $\mathcal{F}$. Note that agent x prefers y to being unmatched if $y \in P(x)$.

A matching $\mathcal{F}$ is said to be *stable* if there exists no *blocking family*, that is a triple $(m, w, d) \notin \mathcal{F}$ such that m prefers w to $\mathcal{F}(m)$, w prefers d to $\mathcal{F}(w)$ and d prefers m to $\mathcal{F}(d)$.

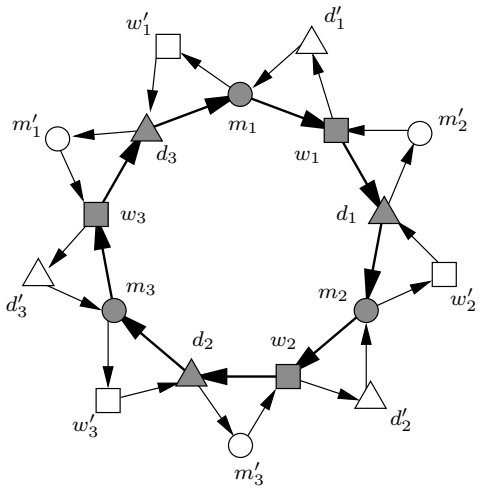
We define the underlying directed graph $D_I = (V, A)$ of an instance I of cyclic 3DSMI as follows. The vertices of D_I correspond to the agents, so $V(D_I) = M \cup W \cup D$, and we have an arc (x, y) in D_I if $y \in P(x)$. This type of directed graph where $A(D_I) \subseteq (M \times W) \cup (W \times D) \cup (D \times M)$ is called a *tripartite cyclic digraph*. Therefore, a matching of I corresponds to a disjoint packing of directed 3-cycles in D_I.

An unsolvable instance of cyclic 3DSMI

We give an instance of cyclic 3DSMI with $n = 6$, denoted by $R6$, where no stable matching exists.

Example 1. *The preference lists and underlying graph of $R6$ are as shown below. Here, the thickness of arrows correspond to preferences.*

$m_1 : w_1, w_1'$	$w_1 : d_1, d_1'$	$d_1 : m_2, m_2'$
$m_2 : w_2, w_2'$	$w_2 : d_2, d_2'$	$d_2 : m_3, m_3'$
$m_3 : w_3, w_3'$	$w_3 : d_3, d_3'$	$d_3 : m_1, m_1'$
$m_1' : w_3$	$w_1' : d_3$	$d_1' : m_1$
$m_2' : w_1$	$w_2' : d_1$	$d_2' : m_2$
$m_3' : w_2$	$w_3' : d_2$	$d_3' : m_3$



We refer to the agents $\{m_i, w_i, d_i : 1 \leq i \leq 3\} = I$ as the *inner agents* of $R6$ and the agents $\{m_i', w_i', d_i' : 1 \leq i \leq 3\} = O$ as the *outer agents* of $R6$.

Lemma 1. *The instance $R6$ of cyclic 3DSMI admits no stable matching.*

Proof. By inspection of the underlying graph of $R6$, we can observe that the only acceptable families are of the form (m_i, w_i', d_{i-1}), (m_i, w_i, d_i') and (m_i', w_{i-1}, d_{i-1}), so that any acceptable family contains exactly two inner agents. It is clear that for any matching $\mathcal{F}$, it must be the case that at least one inner agent is unmatched in $\mathcal{F}$. By the symmetry of the instance we may suppose without loss of generality that the inner agent m_1 is unmatched in $\mathcal{F}$. Then, the family (m_1, w_1', d_3) is a blocking family for $\mathcal{F}$. □

We note that the 9 acceptable families of $R6$ have a natural cyclic order, the same order that the directed 9-cycle has which is formed by the 9 inner agents in the underlying graph, such that if an acceptable family is not in a stable matching $\mathcal{F}$ then the successor family must be in $\mathcal{F}$. For example, if $(m_1, w_1, d_1') \notin \mathcal{F}$ then $(m_2', w_1, d_1) \in \mathcal{F}$, since (m_1, w_1, d_1') would be blocking otherwise. This argument gives an alternative proof for the above Lemma.

The instance created by removing the inner agent m_1 from $R6$, denoted by $R6 \setminus m_1$, becomes solvable, since $\mathcal{F}^* = \{(m_2', w_1, d_1), (m_2, w_2, d_2'), (m_3, w_3', d_2), (m_1', w_3, d_3)\}$ is a stable

matching for $R6 \setminus m_1$. In fact, $\mathcal{F}^*$ is the unique stable matching for $R6 \setminus m_1$, so we denote it by $\mathcal{F}_{R6\setminus m_1}$. This is because in $R6 \setminus m_1$ we have 7 acceptable families in a row with the property discussed above: if an acceptable family is not in a stable matching $\mathcal{F}$ then the subsequent family must be in $\mathcal{F}$. We state this claim formally below; its proof follows from the symmetry of the instance.

Lemma 2. *Let a_i be an inner agent of $R6$. Then, $R6 \setminus a_i$ admits a unique stable matching, denoted by $\mathcal{F}_{R6\setminus a_i}$.*

The instance $R6$ will also be of use to us as a gadget in the NP-completeness proofs of the subsequent sections.

The NP-completeness proof

In [15], Manlove et al. proved that determining if an instance of SMTI admits a complete stable matching is NP-complete, even if the ties appear only on the women's side, and each woman's preference list is either strictly ordered or consists entirely of a tie of size two (these conditions holding simultaneously).

We refer to the MAX SMTI problem under the above restrictions as Restricted SMTI. The underlying graph $G = (A \cup B, E)$ of a Restricted SMTI instance is such that the set $A = \{a_1, a_2, \ldots, a_n\}$ consists of men a_i, all of whom have strictly ordered preference lists, while the set B of women can be partitioned into two sets $B_1 \cup B_2 = \{b_1, \ldots, b_{n_1}\} \cup \{b_1^T, \ldots, b_{n_2}^T\}$ where $n_1 + n_2 = n$, each woman $b_j \in B_1$ has a strictly ordered preference list, and each woman $b_j^T \in B_2$ has a preference list consisting solely of a tie of length 2. We denote a woman who can either be a member of B_1 or B_2 by $b_i^{(T)}$.

In the remainder of this section we describe a polynomial-time reduction from Restricted SMTI to cyclic 3DSMI. Let I be an instance of Restricted SMTI with the underlying graph $G = (A \cup B, E)$. We construct an instance I' of cyclic 3DSMI with sets M, W, and D of men, women, and dogs as follows.

The sets of men and women of I' are created in direct correspondence to the men and women in I, so let $M = \{m_1, \ldots, m_n\}$ and $W = W_1 \cup W_2 = \{w_1, \ldots, w_{n_1}\} \cup \{w_1^T, \ldots, w_{n_2}^T\}$. The set of dogs of I' consists of two parts $D_1 \cup D_2 = D$, defined by creating a dog $d_{j,i}$ in D_1 if $a_i \in P(b_j)$, and creating a dog d_j^T in D_2 if $b_j^T \in B_2$.

Let us now describe the construction of the strictly ordered preference lists of I'. We let $P(x)[l]$ denote the lth entry in agent x's preference list, and a tie in the preference list of an agent is indicated by parentheses. The preference lists of I' are defined by the following cases:

1. If $P(a_i)[l] = b_j^{(T)}$ then let $P(m_i)[l] = w_j^{(T)}$ ($1 \leq l \leq r$, where r is the length of a_i's list).
2. If $P(b_j)[l] = a_i$ then let $P(w_j)[l] = d_{j,i}$ and $P(d_{j,i}) = m_i$ ($1 \leq l \leq r$, where r is the length of b_j's list).
3. If $P(b_j^T) = (a_p, a_q)$ then let $P(w_j^T) = d_j^T$ and $P(d_j^T) = m_p m_q$ (in arbitrary order).

This is the *proper part* of the instance. Next we construct the *additional part* of the instance by creating $n = |M|$ copies of $R6$, such that the t-th copy of $R6$ consists of inner agents $\{m_{t_i}, w_{t_i}, d_{t_i} : 1 \leq i \leq 3\}$ and outer agents $\{m'_{t_i}, w'_{t_i}, d'_{t_i} : 1 \leq i \leq 3\}$ with preference lists as described in Example 1. We add these n copies of $R6$ to the instance in the following way. In the t-th added copy of $R6$, denoted by $R6_t$, replace the inner agent m_{t_1} in $R6_t$ with man $m_t \in M$ by replacing each occurrence of m_{t_1} in the preference lists of each agent in $R6_t$ with m_t. Also, let m_{t_1}'s acceptable partners in $R6_t$, namely w_{t_1} and w'_{t_1} be appended

in this order to the end of m_t's list. The final preference list of man m_t along with $R6_t$ is shown below. The portion of m_t's preference list consisting of women from the proper part of the instance is denoted by P_t.

$$
\begin{array}{rcl}
m_t & : & P_t\ w_{t_1}\ w'_{t_1} \\
m_{t_2} & : & w_{t_2}\ w'_{t_2} \\
m_{t_3} & : & w_{t_3}\ w'_{t_3} \\
m'_{t_1} & : & w_{t_3} \\
m'_{t_2} & : & w_{t_1} \\
m'_{t_3} & : & w_{t_2}
\end{array}
\qquad
\begin{array}{rcl}
w_{t_1} & : & d_{t_1}\ d'_{t_1} \\
w_{t_2} & : & d_{t_2}\ d'_{t_2} \\
w_{t_3} & : & d_{t_3}\ d'_{t_3} \\
w'_{t_1} & : & d_{t_3} \\
w'_{t_2} & : & d_{t_1} \\
w'_{t_3} & : & d_{t_2}
\end{array}
\qquad
\begin{array}{rcl}
d_{t_1} & : & m_{t_2}\ m'_{t_2} \\
d_{t_2} & : & m_{t_3}\ m'_{t_3} \\
d_{t_3} & : & m_t\ m'_{t_1} \\
d'_{t_1} & : & m_t \\
d'_{t_2} & : & m_{t_2} \\
d'_{t_3} & : & m_{t_3}
\end{array}
$$

This ends the reduction, which plainly can be computed in polynomial time. Now, we prove that there is a one-to-one correspondence between the complete stable matchings in I and the stable matchings in I'.

First we show that there is a one-to-one correspondence between the matchings of I and the matchings in the proper part of I'. This comes from the natural one-to-one correspondence between the edges of I and the families in the proper part of I'. More precisely, if $\mathcal{M}$ is a matching in I, then the corresponding matching $\mathcal{F}_p$ in the proper part of I is created as follows: $(a_i, b_j) \in \mathcal{M} \iff (m_i, w_j, d_{j,i}) \in \mathcal{F}_p$ and $(a_i, b_j^T) \in \mathcal{M} \iff (m_i, w_j^T, d_j^T) \in \mathcal{F}_p$. To prove this, it is enough to observe that two edges in I are disjoint if and only if the two corresponding families in I' are also disjoint. Next, we show that stability is preserved by this correspondence.

Lemma 3. *A matching $\mathcal{M}$ of I is stable if and only if the corresponding matching $\mathcal{F}_p$ in the proper part of I' is stable.*

Proof. It is enough to show that an edge (a_i, b_j) is blocking in I if and only if the corresponding family $(m_i, w_j, d_{j,i})$ is also blocking in I'; and similarly, an edge (a_i, b_j^T) is blocking in I if and only if the corresponding family (m_i, w_j^T, d_j^T) is also blocking in I'.

Suppose first that (a_i, b_j) is blocking in I, which means that a_i is either unmatched or prefers b_j to $\mathcal{M}(a_i)$ and b_j is either unmatched or prefers a_i to $\mathcal{M}(b_j)$. This implies that m_i prefers w_j to $\mathcal{F}_p(m_i)$, w_j prefers $d_{j,i}$ to $\mathcal{M}(w_j)$, and $d_{j,i}$ is unmatched in $\mathcal{F}_p$, i.e. $(m_i, w_j, d_{j,i})$ is blocking in I'. Similarly, if (a_i, b_j^T) is blocking then a_i is either unmatched or prefers b_j^T to $\mathcal{M}(a_i)$ and b_j^T is unmatched in $\mathcal{M}$. This implies that m_i prefers w_j^T to $\mathcal{F}_p(m_i)$, w_j^T and d_j^T are both unmatched in $\mathcal{F}_p$, and hence (m_i, w_j^T, d_j^T) is blocking in I'.

In the other direction, if $(m_i, w_j, d_{j,i})$ is blocking in I', then m_i prefers w_j to $\mathcal{F}_p(m_i)$, w_j prefers $d_{j,i}$ to $\mathcal{F}_p(w_j)$, and $d_{j,i}$ is unmatched in $\mathcal{F}_p$. This implies that a_i is either unmatched or prefers b_j to $\mathcal{M}(a_i)$ and b_j is either unmatched or prefers a_i to $\mathcal{M}(b_j)$, so (a_i, b_j) is blocking in I. Similarly, if (m_i, w_j^T, d_j^T) is blocking in I', then w_j^T and d_j^T are both unmatched in $\mathcal{F}_p$ and m_i prefers w_j^T to $\mathcal{F}_p(m_i)$. This implies that a_i is either unmatched or prefers b_j^T to $\mathcal{M}(a_i)$ and b_j^T is unmatched in $\mathcal{M}$, so (a_i, b_j^T) is blocking in I. □

Furthermore, if the matching $\mathcal{M}$ is complete, then we can enlarge the corresponding matching to the additional part of I' by matching every $R6_t \setminus m_t$ in the unique stable way, so by adding $\mathcal{F}_{R6_t \setminus m_t}$ to $\mathcal{F}_p$ for every t. This leads to the following one-to-one correspondence between the complete stable matchings of I and the stable matching of I'.

Lemma 4. *The instance I admits a complete stable matching $\mathcal{M}$ if and only if the reduced instance I' admits a stable matching $\mathcal{F}$, where $\mathcal{F}$ is the corresponding matching of $\mathcal{M}$.*

Proof. The stability of $\mathcal{M}$ implies that $\mathcal{F}$ is stable in the proper part of I' by Lemma 3. The completeness of $\mathcal{M}$ and Lemma 2 implies that $\mathcal{F}$ is also stable in the additional part of I'.

In the other direction, if $\mathcal{F}$ is stable then every man in M must be matched in a proper family, since otherwise, if a proper man m_t does not have a proper partner in $\mathcal{F}$ then $R6_t$ would contain a blocking family, by Lemma 1. This implies that the corresponding matching $\mathcal{M}$, defined in Lemma 3, is complete. The stability of $\mathcal{M}$ is a consequence of Lemma 3. Finally, we note that the additional part has a unique stable matching, since every $R6_t \setminus a_t$ must be matched in the unique stable way indicated by Lemma 2, which implies the one-to-one correspondence. □

The following Theorem is a direct consequence of Lemma 4.

Theorem 1. *Determining the existence of a stable matching in a given instance of cyclic* 3DSMI *is NP-complete.*

3 Cyclic 3DSM under strong stability is NP-complete

Problem definition

For an instance of cyclic 3DSM, a matching $\mathcal{F}$ is *strongly stable* if there exists no *weakly blocking family*, that is a family $(m, w, d) \notin \mathcal{F}$ such that m prefers w to $\mathcal{F}(m)$ or $w = \mathcal{F}(m)$, w prefers d to $\mathcal{F}(w)$ or $d = \mathcal{F}(w)$, and d prefers m to $\mathcal{F}(d)$ or $m = \mathcal{F}(d)$. We note that in a weakly blocking family at least two members obtain a better partner, since the preference lists are strictly ordered.

An unsolvable instance

We firstly show that, by completing the preference lists of $R6$ in an arbitrary way (i.e. by appending agents not on the lists in an arbitrary order to the tail of the original lists), the resulting instance of cyclic 3DSM, denoted by $\overline{R6}$, does not admit any strongly stable matching. The subinstance $R6$ of $\overline{R6}$ is called the *suitable part* of $\overline{R6}$, the original entries of an agent x in $R6$ are the *suitable partners of* x and the families of $R6$ are called *suitable families*.

Lemma 5. *The instance $\overline{R6}$ of cyclic 3DSM admits no strongly stable matching.*

Proof. Suppose for contradiction that $\mathcal{F}$ is a strongly stable matching. As the 9 inner agents form a 9-cycle in the underlying directed graph, the 9 suitable families have a natural cyclic order. We show that if a suitable family, say (m_1, w_1, d_1') is not in $\mathcal{F}$, then the successor suitable family (m_2', w_1, d_1) must be in $\mathcal{F}$, which would imply a contradiction given that the number of these suitable families is odd. If $(m_1, w_1, d_1') \notin \mathcal{F}$ then $\mathcal{F}(w_1) = d_1$, since otherwise (m_1, w_1, d_1') would be weakly blocking. Similarly, $(m_2', w_1, d_1) \notin \mathcal{F}$ implies $\mathcal{F}(d_1) = m_2$. But this means that $(m_2, w_1, d_1) \in \mathcal{F}$, so (m_2, w_2', d_1) is weakly blocking. □

Recall that $\mathcal{F}_{R6\setminus a_t}$ is the unique stable matching for $R6 \setminus a_t$. Let $\overline{R6} \setminus a_t$ denote the instance created by removing an inner agent a_t from $\overline{R6}$. We denote by $C_{R6\setminus a_t}$ the subset of agents of $\overline{R6} \setminus a_t$ that are covered by $\mathcal{F}_{R6\setminus a_t}$, and by $U_{R6\setminus a_t}$ those who are uncovered by $\mathcal{F}_{R6\setminus a_t}$, respectively.

Lemma 6. *Let a_t be an inner agent of $\overline{R6}$. For every matching $\mathcal{F}^* \supseteq \mathcal{F}_{R6\setminus a_t}$ of $\overline{R6} \setminus a_t$, no suitable family can be weakly blocking, and therefore no agent from $C_{R6\setminus a_t}$ can be involved in a weakly blocking family. For any other matching, at least one suitable family is weakly blocking.*

Proof. It is straightforward to verify that $\mathcal{F}_{R6\setminus a_t}$ is a strongly stable matching for $R6 \setminus a_t$, so no suitable family in $\overline{R6} \setminus a_t$ can weakly block $\mathcal{F}^* \supseteq \mathcal{F}_{R6\setminus a_t}$. Moreover, no agent x of $C_{R6\setminus a_t}$ can be involved in a non-suitable weakly blocking family either, since x has a suitable partner in $\mathcal{F}^*$.

Suppose that $\mathcal{F}'$ is a matching of $\overline{R6} \setminus a_t$ which is not a superset of $\mathcal{F}_{R6\setminus a_t}$. As in the proof of Lemma 5, we use the fact that if a suitable family is not in $\mathcal{F}'$, then the successor suitable family is either in $\mathcal{F}'$ or weakly blocking. Therefore, if we do not include four from the seven suitable families of $\overline{R6} \setminus a_t$ in a matching then one of them would be weakly blocking. □

The NP-completeness proof

The reduction we describe in this section again begins with an instance of Restricted SMTI, only we assume without loss of generality the role of the men and women of the instance to be "reversed". To be precise, we assume a given instance of Restricted SMTI I that its vertex set $((A_1 \cup A_2) \cup B)$ consists of a set $A_1 = \{a_1, a_2, \ldots, a_{n_1}\}$ of men with strictly ordered preference lists, and $A_2 = \{a_1^T, a_2^T, \ldots, a_{n_2}^T\}$ of men with preference lists consisting of a single tie of length 2, and $n_1 + n_2 = n$. The set $B = \{b_1, b_2, \ldots, b_n\}$ consists entirely of women with strictly ordered preference lists.

Given an instance I of Restricted SMTI as defined above, we create an instance I' of cyclic 3DSM. First we create a *proper instance* I'_p of cyclic 3DSMI as a subinstance of I' with agents $M_p \cup W_p \cup D_p$ in the following way.

First we create a set W_p of n women $\{w_1, w_2, \ldots, w_n\}$ such that the preference list of woman w_j is a single entry, dog $d_j \in D_p$. The preference list of d_j is such that if $P(b_j)[l] = a_i$, then $P(d_j)[l] = m_i$, otherwise if $P(b_j)[l] = a_i^T$, then $P(d_j)[l] = m'_{i,j}$ for $1 \leq l \leq r$, where r is the length of b_j's list. So the preference list of dog d_j is essentially the "same" as that of woman b_j, only with men in M_p rather than A.

For each man $a_i \in A_1$, create a man $m_i \in M_p$, such that if $P(a_i)[l] = b_j$, then let $P(m_i)[l] = w_j$ for $1 \leq l \leq r$, where r is the length of a_i's list. So the preference list of man m_i is essentially the "same" as that of man a_i. For each man $a_i^T \in A_2$, with a preference list consisting of a single tie of length two, say (b_r, b_s), we create five men $m_i^T, m'_{i,r}, m''_{i,r}, m'_{i,s}, m''_{i,s}$, four women $w'_{i,r}, w''_{i,r}, w'_{i,s}, w''_{i,s}$ and four dogs $d'_{i,r}, d''_{i,r}, d'_{i,s}, d''_{i,s}$ where the preference list of m_i^T contains $w'_{i,r}$ and $w'_{i,s}$ in an arbitrary order, and the other preference lists are as shown below.

$m'_{i,r}$:	$w'_{i,r}$	w_r	$w'_{i,r}$:	$d'_{i,r}$	$d''_{i,r}$	$d'_{i,r}$:	$m''_{i,r}$ m_i^T
$m''_{i,r}$:	$w''_{i,r}$		$w''_{i,r}$:	$d''_{i,r}$	$d'_{i,r}$	$d''_{i,r}$:	$m'_{i,r}$ $m''_{i,r}$
$m'_{i,s}$:	$w'_{i,s}$	w_s	$w'_{i,s}$:	$d'_{i,s}$	$d''_{i,s}$	$d'_{i,s}$:	$m''_{i,s}$ m_i^T
$m''_{i,s}$:	$w''_{i,s}$		$w''_{i,s}$:	$d''_{i,s}$	$d'_{i,s}$	$d''_{i,s}$:	$m'_{i,s}$ $m''_{i,s}$

We also add these agents to M_p, W_p and D_p, respectively. Note that in I'_p every set of agents has the same cardinality: $n_p = |M_p| = |W_p| = |D_p| = n + 4n_2$. The notions of *proper agent, proper partner* and *proper family* are defined in the obvious way.

The *additional part* of instance I' contains three subinstances. The *suitable part* of I' is the disjoint union of $3n_p$ copies of $R6$, such that the ith copy of $R6$, denoted $R6_i$, incorporates the ith agent of I'_p, as described in the previous reduction in the proof of Theorem 1 (we omit the full description of this process again). The new agents are referred to as *additional agents*.

Let $\mathcal{F}_s = \cup_{i\in\{1,\ldots 3n_p\}}\mathcal{F}_{R6_i\setminus a_i}$ be the so-called *suitable matching* of the additional part, where a_i is the proper agent of $R6_i$. We call the set $C = \cup_{i\in\{1,\ldots 3n_p\}}C_{R6_i\setminus a_i}$ *covered additional agents*, as these additional agents are covered by $\mathcal{F}_s$, and we call the set $U = \cup_{i\in\{1,\ldots 3n_p\}}U_{R6_i\setminus a_i}$ *uncovered additional agents*, as these additional agents are not covered by $\mathcal{F}_s$.

The *fitting part* of I' is constructed on U as follows. Note that U has equal numbers of men, women and dogs. The fitting part consists of disjoint families that covers U, so that every agent has exactly one agent in his/her/its list, i.e. the fitting part is a complete matching of U, denoted by $\mathcal{F}_f$.

Finally, the *dummy part* is obtained by an arbitrary extension of the preference lists, so that by putting together the four subinstances, the proper and the three additional parts, we get the complete instance I'. The preferences of the agents over the partners in different parts respect the order in which we defined these parts: the list of a proper agent contains the proper partners first, then the suitable partners, and finally the dummy partners; the list of a covered additional agent contains the suitable partners first, then the dummy partners; the list of an uncovered additional agent contains the suitable partners first, then the fitting partner, and finally the dummy partners.

First we show that there is a one-to-one correspondence between the complete stable matchings of I and the complete strongly stable matchings of I'_p. The stability is preserved via the following one-to-one correspondence between the complete matchings of I and complete matchings of I':

$$(a_i, b_j) \in \mathcal{M} \iff (m_i, w_j, d_j) \in \mathcal{F}_p$$

$$(a_i^T, b_s) \in \mathcal{M} \iff (m_i^T, w'_{i,s}, d'_{i,s}), (m''_{i,s}, w''_{i,s}, d''_{i,s}), (m'_{i,s}, w_s, d_s) \in \mathcal{F}_p$$

$$(a_i^T, b_s) \notin \mathcal{M} \iff (m'_{i,s}, w'_{i,s}, d''_{i,s}), (m''_{i,s}, w''_{i,s}, d'_{i,s}) \in \mathcal{F}_p$$

Lemma 7. *A complete matching $\mathcal{M}$ of I is stable if and only if the corresponding complete matching $\mathcal{F}_p$ of I'_p is strongly stable.*

Proof. As a man a_i^T cannot belong to a blocking pair in I, it may be verified that his corresponding copy m_i^T cannot belong to a weakly blocking family in I_p either. Therefore, it is enough to show that a pair (a_i, b_j) is blocking for $\mathcal{M}$ if and only if the corresponding family (m_i, w_j, d_j) is blocking for $\mathcal{F}_p$. But this is obvious, because the preference lists of a_i and m_i are essentially the same, and the preference lists of b_j and d_j are also essentially the same. □

Now, given a matching $\mathcal{M}$ of I let us create the corresponding matching $\mathcal{F}$ of I' by adding $\mathcal{F}_s$ and $\mathcal{F}_f$ to $\mathcal{F}_p$, so $\mathcal{F} = \mathcal{F}_p \cup \mathcal{F}_s \cup \mathcal{F}_f$.

Lemma 8. *The instance I admits a complete stable matching $\mathcal{M}$ if and only if the reduced instance I' admits a strongly stable matching $\mathcal{F}$, where $\mathcal{F}$ is the corresponding matching of $\mathcal{M}$.*

Proof. Suppose that we have a complete stable matching $\mathcal{M}$ of I, and $\mathcal{F}$ is the corresponding matching in I'. Lemma 7 implies that every proper agent has a proper partner in $\mathcal{F}$ and no proper family is weakly blocking. Therefore, no proper agent can be involved in any weakly blocking family either. By construction of $\mathcal{F}_s$, every covered additional agent has a suitable partner in $\mathcal{F}$ and by Lemma 6, no suitable family is weakly blocking. Therefore, no such agent can be part of any weakly blocking family. Finally, every uncovered additional agent has a fitting partner in $\mathcal{F}$, so these agent cannot form a weakly blocking family either, since an uncovered additional agent prefers only suitable partners to fitting partners, which cannot be involved in a weakly blocking family. Hence $\mathcal{F}$ is strongly stable.

In the other direction, suppose that $\mathcal{F}$ is a strongly stable matching of I'. Every proper agent must have a proper partner, since otherwise if a_t had no proper partner in $\mathcal{F}$, then $\overline{R6}_t$ would contain a suitable weakly blocking family by Lemma 5. So the corresponding matching $\mathcal{M}$ in I is complete. The stability of $\mathcal{M}$ is a consequence of Lemma 7. Finally, we note that the additional agents must be matched in the unique strongly stable way in $\mathcal{F}$, namely, the covered additional agents must be covered by matching $\mathcal{F}_s$ by Lemma 6, and the uncovered additional agents must be covered by $\mathcal{F}_f$, since otherwise a fitting family would weakly block $\mathcal{F}$. Therefore, we have a one-to-one correspondence as was claimed. □

Theorem 2. *Determining the existence of a strongly stable matching in a given instance of cyclic* 3DSM *is NP-complete.*

4 Stable exchanges with restrictions

Problem definition

Given a simple digraph $D = (V, A)$, where V is the set of agents, suppose that each agent has exactly one indivisible good, and $(i, j) \in A$ if the good of agent j is suitable for agent i. An *exchange* is a permutation π of V such that, for each $i \in V$, $i \neq \pi(i)$ implies $(i, \pi(i)) \in A$. Alternatively, an exchange can be considered as a disjoint packing of directed cycles in D.

Let each agent have strict preferences over the goods, that are suitable for him. These orderings can be represented by preference lists. In an exchange π, the agent i receives the good of his *successor*, $\pi(i)$; therefore the agent i prefers an exchange π to another exchange σ if he prefers $\pi(i)$ to $\sigma(i)$. An exchange π is *stable* if there is no *blocking coalition* B, i.e. a set B of agents and a permutation σ of B where every agent $i \in B$ prefers σ to π. An exchange is *strongly stable* is there exists no *weakly blocking coalition* B with a permutation σ of B where for every agent $i \in B$, either $\sigma(i) = \pi(i)$ or i prefers σ to π, and $\sigma(i) \neq \pi(i)$ for at least one agent $i \in B$.

Complexity results about stable exchanges

Shapley and Scarf [25] showed that the stable exchange problem is always solvable and a stable exchange can be found in polynomial time by the Top Trading Cycle (TTC) algorithm, proposed by Gale. Moreover, Roth and Postlewaite [18] proved that the exchange obtained by the TTC algorithm is strongly stable and this is the only such solution. We note that they considered this problem as a so-called *houseswapping game*, where a *core element* corresponds to a stable solution. (For further details about these connections with Game Theory, see [3].)

In some applications the length of the possible cycles is bounded by some constant l. In this case we consider an *l-way exchange problem*. Furthermore, the size of the possible blocking coalitions can also be restricted. We say that an exchange is *b-way stable* if there exists no blocking coalition of size at most b. Because of some applications, the most relevant problems are for constants 2 and 3. Henceforth we also refer to "2-way" as "pairwise" in the context of cycle lengths and blocking coalitions sizes. We remark that if $b = l$ then a stable exchange corresponds to a core-solution of some related NTU-game, because the possible coalitions, those that can form and those that can block, are the same (see [3] for details).

For $l = b = 2$, the pairwise stable pairwise exchange problem is in fact, equivalent to the *stable roommates problem*. Therefore, a stable solution may not exist [9], but there is a polynomial-time algorithm that finds a stable solution if one does exist [11] or reports that none exists. For $l = b = 3$, the 3-way stable 3-way exchange problem is NP-hard, even for three-sided directed graphs, as is stated by the following theorem.

Theorem 3. *The 3-way stable 3-way exchange problem for tripartite directed graphs is equivalent to the cyclic 3DSMI problem, and is therefore NP-complete.*

Finally, we note that Irving [12] proved recently that the stable pairwise exchange and the 3-way stable pairwise exchange problems are NP-hard. The pairwise stable 3-way exchange problem is open. This particular problem can be a relevant regarding the application of kidney exchanges, next described.

Kidney exchange problem

Living donation is the most effective treatment that is currently known for kidney failure. However a patient who requires a transplant may have a willing donor who cannot donate to them for immunological reasons. So these incompatible patient-donor pairs may want to exchange kidneys with other pairs. Kidney exchange programs have already been established in several countries such as the Netherlands [13] and the USA [20].

In most of the current programs the goal is to maximise the number of patients that receive a suitable kidney in the exchange [21, 22, 23, 1] by regarding only the eligibility of the grafts. Some more sophisticated variants consider also the difference between suitable kidneys. Sometimes the "total benefit" is maximised [24], whilst other models [19, 6, 7, 4] require first the stability of the solution under various criteria.

The length of the cycles in the exchanges is bounded in the current programs, because all operations along a cycle have to be carried out simultaneously. Most programs allow only pairwise exchanges. But sometimes 3-way exchanges are also possible, like in the New England Program [16] and in the National Matching Scheme of the UK [27][2]. In these kind of applications, if one considers stability as the first priority of the solution, then we obtain a 3-way stable 3-way exchange problem, where the incompatible patient-donor pairs are the agents and their preferences are determined according to the special parameters of the suitable kidneys.

Finally, we remark that although the induced digraph of a real kidney exchange instance may have special properties (see e.g. [22] about the effect of the blood-types on the digraph) the problem remains hard, even for realistic situations. For example, if we have three sets of patient-donor pairs with blood types O-A, A-B and B-O, then the digraph may appear to be tripartite. But this particular case of the 3-way stable 3-way exchange problem is also hard by Theorem 3.

5 Further questions

For cyclic 3DSMI, the smallest instance that admits no stable matching given here satisfies $n = 6$. Is there an even smaller counterexample? In the case of strong stability, we are aware of instances of cyclic 3DSM for $n = 4$ that admit no strongly stable matching.

The main questions that remain unsolved are (i) whether there exists an instance of cyclic 3DSM that admits no stable matching, and (ii) whether there is a polynomial-time algorithm to find such a matching or report that none exists, given an instance of cyclic 3DSM.

[2] 3-way exchanges may be also allowed in the national program of the USA (as it is declared to be a goal of the system in the future in the Proposal for National Paired Donation Program [28]).

References

[1] D. J. Abraham, A. Blum, and T. Sandholm. Clearing algorithms for barter-exchange markets: Enabling nationwide kidney exchanges. *In Proceedings of ACM-EC 2007: the Eighth ACM Conference on Electronic Commerce*, 2007.

[2] A. Alkan. Non-existence of stable threesome matchings. *Mathematical Social Sciences*, 16:207–209, 1988.

[3] P. Biró. *The stable matching problem and its generalizations: an algorithmic and game theoretical approach*. PhD thesis, Budapest University of Technology and Economics, 2007.

[4] P. Biró and K. Cechlárová. Inapproximability of the kidney exchange problem. *Inf. Process. Lett.*, 101(5):199–202, 2007.

[5] E. Boros, V. Gurvich, S. Jaslar, and D. Krasner. Stable matchings in three-sided systems with cyclic preferences. *Discrete Mathematics*, 289:1–10, 2004.

[6] K. Cechlárová, T. Fleiner, and D. F. Manlove. The kidney exchange game. *Proc. SOR'05*, Eds. L. Zadik-Stirn, S. Drobne:77–83, 2005.

[7] K. Cechlárová and V. Lacko. The kidney exchange game: How hard is to find a donor? *IM Preprint*, 4/2006, 2006.

[8] K. Eriksson, J. Sjöstrand, and P. Strimling. Three-dimensional stable matching with cyclic preferences. *Math. Social Sci.*, 52(1):77–87, 2006.

[9] D. Gale and L. S. Shapley. College Admissions and the Stability of Marriage. *Amer. Math. Monthly*, 69(1):9–15, 1962.

[10] Chien-Chung Huang. Two's company, three's a crowd: stable family and threesome roommate problems. In *Algorithms—ESA 2007*, volume 4698 of *Lecture Notes in Comput. Sci.*, pages 558–569. Springer, Berlin, 2007.

[11] R.W. Irving. An efficient algorithm for the "stable roommates" problem. *Journal of Algorithms*, 6:577–595, 1985.

[12] R.W. Irving. The cycle-roommates problem: a hard case of kidney exchange. *Information Processing Letters*, 103:1–4, 2007.

[13] K. M. Keizer, M. de Klerk, B. J. J. M. Haase-Kromwijk, and W. Weimar. The Dutch algorithm for allocation in living donor kidney exchange. *Transplantation Proceedings*, 37:589–591, 2005.

[14] D.E. Knuth. *Mariages Stables*. Les Presses de L'Université de Montréal, 1976.

[15] D.F. Manlove, R.W. Irving, K. Iwama, S. Miyazaki, and Y. Morita. Hard variants of stable marriage. *Theoretical Computer Science*, 276(1-2):261–279, 2002.

[16] New England Program for Kidney Exchange. *http://www.nepke.org*.

[17] C. Ng and D. S. Hirschberg. Three-dimensional stable matching problems. *SIAM J. Discrete Math.*, 4(2):245–252, 1991.

[18] A. E. Roth and A. Postlewaite. Weak versus strong domination in a market with indivisible goods. *J. Math. Econom.*, 4(2):131–137, 1977.

[19] A. E. Roth, T. Sönmez, and U. M. Ünver. Kidney exchange. *J. Econom. Theory*, 119:457–488, 2004.

[20] A. E. Roth, T. Sönmez, and U. M. Ünver. A kidney exchange clearinghouse in New England. *American Economic Review, Papers and Proceedings*, 95(2):376–380, 2005.

[21] A. E. Roth, T. Sönmez, and U. M. Ünver. Pairwise kidney exchange. *J. Econom. Theory*, 125(2):151–188, 2005.

[22] A. E. Roth, T. Sönmez, and U. M. Ünver. Coincidence of wants in markets with compatibility based preferences. *American Economic Review*, 97(3):828–851, 2007.

[23] S. L. Saidman, A. E. Roth, T. Sönmez, U. M. Ünver, and S. L. Delmonico. Increasing the opportunity of live kidney donation by matching for two and three way exchanges. *Transplantation*, 81(5):773–782, 2006.

[24] S. L. Segev, S. E. Gentry, D. S. Warren, B. Reeb, and R. A. Montgomery. Kidney paired donation and optimizing the use of live donor organs. *J. Am. Med. Assoc.*, 293:1883–1890, 2005.

[25] L. S. Shapley and H. E. Scarf. On cores and indivisibility. *J. Math. Econom.*, 1(1):23–37, 1974.

[26] A. Subramanian. A new approach to stable matching problems. *SIAM Journal on Computing*, 23(4):671–700, 1994.

[27] UK Transplant. *http://www.uktransplant.org.uk*.

[28] United Network for Organ Sharing. *http://www.unos.org*.

Péter Biró and Eric McDermid
Department of Computing Science
University of Glasgow
Glasgow G12 8QQ, UK
Email: {pbiro,mcdermid}@dcs.gla.ac.uk

Equilibria in Social Belief Removal[1]

Richard Booth and Thomas Meyer

Abstract

In studies of multi-agent interaction, especially in game theory, the notion of *equilibrium* often plays a prominent role. A typical scenario for the *belief merging* problem is one in which several agents pool their beliefs together to form a consistent "group" picture of the world. The aim of this paper is to define and study new notions of equilibria in belief merging. To do so, we assume the agents arrive at consistency via the use of a *social belief removal* function, in which each agent, using his own *individual* removal function, removes some belief from his stock of beliefs. We examine several notions of equilibria in this setting, assuming a general framework for individual belief removal due to Booth et al. We look at their inter-relations as well as prove their existence or otherwise. We also show how our equilibria can be seen as a generalisation of the idea of taking maximal consistent subsets of agents.

1 Introduction

The problem of multi-agent belief merging has received a lot of attention in KR in recent years [13, 14, 5]. The problem occurs when several agents each have their own beliefs, and want to combine or pool them into a consistent "group" picture of the world. A problem arises when two or more agents have conflicting beliefs. Then such conflicts need to be smoothed out. In studies of multi-agent interaction the notion of *equilibrium* often plays a prominent role (most famously in [18]). It would therefore seem natural to investigate such notions in belief merging. The purpose of this paper is to define and study some possible notions of equilibria in a belief merging setting.

To enable a clear formulation of such notions, we will employ the approach to merging advocated in [5] and inspired by the contraction+expansion approach to belief revision [9, 16], in which the merging operation is explicitly broken down into two sub-operations. In the first stage, the agents each modify their own beliefs in such a way as to make them jointly consistent. This is called *social contraction* in [5]. In the second, trivial, stage, the beliefs thus obtained are conjoined. In this approach, the crucial question becomes "how do the agents modify their beliefs in the first stage?" In this paper we assume agents do so by *removing* some sentence from their stock of beliefs. More precisely we associate to each agent i its very own *individual* removal function $\divideontimes_i$ which computes the result of removing any given sentence. A *social belief removal* function is then a function which, given a profile of individual removal functions as input, returns a (consistent) profile consisting of the results of each agent's removal. The central question studied in this paper is *"when can the outcome of a social removal function be said to be in equilibrium?"*.

How can we express the idea of equilibrium in social removal? As our starting point we would like to propose the following general principle for multi-agent interaction:

Principle of Equilibrium
Each agent simultaneously makes the appropriate *response to what all the other agents do.*

It remains to formalise what "appropriate" means. In the theory of *strategic games* (see, e.g., [20] as well as Section 6 of the present paper) agents are assumed to have their own preferences over the set of all outcomes. Then a *Nash equilibrium* [18] is a profile consisting of each agent's selected action, in which no agent can achieve a more preferred outcome by changing his action, given the actions of the other agents are held fixed. Hence in this setting "appropriate" may be equated with

[1] A longer version of this paper (including the more important proofs) appears in the Proceedings of the 11^{th} International Conference on Principles of Knowledge Representation and Reasoning (KR 2008).

"best" in a precise sense. We will see that the framework of social belief removal offers up new and interesting ways of formalising what "appropriate" might mean.

Of course the explicit introduction of individuals' removal functions raises the question of what kind of belief removal function we should assume is being used. Do agents use AGM contraction [1], or severe withdrawal [22], or perhaps a belief liberation function [6]? Luckily there exists a general family, called *basic removal* [7] which contains *all* these families and more besides. Thus we find it convenient to use this family as a basis.

The plan of the paper is as follows. In Section 2 we set up the framework of social removal functions. Then in Section 3 we focus on the agents' individual removal functions, reviewing some results about basic removal functions and giving some concrete examples of such functions. In Section 4 we introduce our first equilibrium notion, that of a *removal equilibrium*, and examine its compatibility with some plausible minimal change properties, before proving the existence of such equilibria for arbitrary basic removal profiles in Section 5. We also briefly look at the notion of *perfect removal equilibria*. In Section 6 we move on to *entrenchment equilibria*, which can be thought of as Nash equilibria of the strategic game where agent preferences over outcomes are derived from their entrenchment orderings, and examine their relationship with removal equilibria. We also suggest a possible refinement of this idea, the *strong entrenchment equilibrium*. In Section 7 we show how our equilibria can be thought of as generalising the idea of taking maximal consistent subsets of agent. We finish with a concluding section.

Preliminaries: We work in a finitely-generated propositional language L. Classical logical consequence and logical equivalence are denoted by $\vdash$ and $\equiv$ respectively. W denotes the set of possible worlds/interpretations for L. Given $\theta \in L$, we denote the set of worlds in which θ is true by $[\theta]$. The set of non-tautologous sentences in L is denoted by L_*. We will usually talk of belief *sets*, but assume a belief set is always represented by a single sentence standing for its set of logical consequences. We assume a set of *agents* $\mathbb{A} = \{1, \ldots, n\}$. A *belief profile* is any n-tuple of belief sets. Given two belief profiles we shall write $(\phi_i)_{i\in\mathbb{A}} \equiv (\phi'_i)_{i\in\mathbb{A}}$ iff $\phi_i \equiv \phi'_i$ for all i, and write $(\phi_i)_{i\in\mathbb{A}} \equiv_\wedge (\phi'_i)_{i\in\mathbb{A}}$ iff $\bigwedge_{i\in\mathbb{A}} \phi_i \equiv \bigwedge_{i\in\mathbb{A}} \phi'_i$. Clearly we have $\equiv \subseteq \equiv_\wedge$ for belief profiles. We say the belief profile is consistent iff the conjunction of its elements is consistent.

2 Social belief removal

As we said above, we assume each agent $i \in \mathbb{A}$ comes equipped with its own *removal function* $\divideontimes_i$, which tells it how to remove any given sentence from its belief set. In this paper we view $\divideontimes_i$ as a *unary function* on the set L_* of non-tautologous sentences, i.e., agents are never required to remove $\top$. The result of removing $\lambda \in L_*$ from i's belief set is denoted by $\divideontimes_i(\lambda)$. We assume i's *initial* belief set can always be recaptured from $\divideontimes_i$ alone by just removing the contradiction, i.e., i's initial belief set is $\divideontimes_i(\bot)$. We call any n-tuple $(\divideontimes_i)_{i\in\mathbb{A}}$ of removal functions a *removal profile*.

Definition 1 *A* social removal function $\mathbf{F}$ *(relative to $\mathbb{A}$) is any function which takes as input any removal profile $(\divideontimes_i)_{i\in\mathbb{A}}$ and outputs a consistent belief profile $\mathbf{F}((\divideontimes_i)_{i\in\mathbb{A}}) = (\phi_i)_{i\in\mathbb{A}}$ such that, for each $i \in \mathbb{A}$, there exists $\lambda_i \in L_*$ such that $\phi_i \equiv \divideontimes_i(\lambda_i)$.*

Each social removal function yields a merging operator for removal profiles – we just take the conjunction $\bigwedge_{i\in\mathbb{A}} \phi_i$ of the agents' new belief profile. However in this paper our main interest will be in the profile itself.

The above definition differs from Booth's social contraction in two main ways. First, here we *explicitly* associate from the outset an individual removal function to each i, whereas this was only implicit in [5]. More importantly, unlike in social contraction, we will allow agents to use removal functions which don't necessarily satisfy the Inclusion property, i.e., removing a sentence *may* lead to new beliefs entering i's belief set. As is argued in [6], this situation can arise quite naturally. This motivates the use of the term social *removal* rather than social *contraction*.

What properties might we expect from a social removal function $\mathbf{F}$? Throughout the paper we will mention various postulates for $\mathbf{F}$, but to begin with the following two properties have – on the face of it – a strong appeal from a "minimal change" viewpoint:

(FVac) If $(\divideontimes_i(\bot))_{i\in\mathbb{A}}$ is consistent then $\mathbf{F}((\divideontimes_i)_{i\in\mathbb{A}}) \equiv (\divideontimes_i(\bot))_{i\in\mathbb{A}}$

(FVac$_\wedge$) If $(\divideontimes_i(\bot))_{i\in\mathbb{A}}$ is consistent then $\mathbf{F}((\divideontimes_i)_{i\in\mathbb{A}}) \equiv_\wedge (\divideontimes_i(\bot))_{i\in\mathbb{A}}$

Both these rules deal with the case the initial belief sets of the agents are already jointly consistent. **(FVac)** says that in this case the agents' beliefs should remain unchanged. Although intuitively appealing, we will later have grounds for believing this rule is a touch too strong (specifically in contexts where the agents' individual removal functions might not adhere to the Vacuity rule – see next section). Rule **(FVac$_\wedge$)** is weaker. It requires only that the result should be conjunction-equivalent to the profile of the agents' initial belief sets.

3 Basic and hyperregular removal

What properties should be assumed of the individual removal functions $\divideontimes_i$? We will assume agents always use *basic* removal [7].

Definition 2 *A function* $\divideontimes : L_* \to L$ *is a* basic removal function *iff it satisfies the following rules:*

($\divideontimes$1) $\divideontimes(\lambda) \nvdash \lambda$

($\divideontimes$2) *If* $\lambda_1 \equiv \lambda_2$ *then* $\divideontimes(\lambda_1) \equiv \divideontimes(\lambda_2)$

($\divideontimes$3) *If* $\divideontimes(\chi \wedge \lambda) \vdash \chi$ *then* $\divideontimes(\chi \wedge \lambda \wedge \psi) \vdash \chi$

($\divideontimes$4) *If* $\divideontimes(\chi \wedge \lambda) \vdash \chi$ *then* $\divideontimes(\chi \wedge \lambda) \vdash \divideontimes(\lambda)$

($\divideontimes$5) $\divideontimes(\chi \wedge \lambda) \vdash \divideontimes(\chi) \vee \divideontimes(\lambda)$

($\divideontimes$6) *If* $\divideontimes(\chi \wedge \lambda) \nvdash \lambda$ *then* $\divideontimes(\lambda) \vdash \divideontimes(\chi \wedge \lambda)$

All these rules are familiar from the literature on belief removal. Rule **($\divideontimes$1)** is the Success postulate which says the sentence to be removed is no longer implied by the new belief set, while **($\divideontimes$2)** is a syntax-irrelevance property. Rule **($\divideontimes$3)** is sometimes known as Conjunctive Trisection [11, 21]. It says if χ is believed after removing the conjunction $\chi \wedge \lambda$, then it should also be believed when removing the longer conjunction $\chi \wedge \lambda \wedge \psi$. Rule **($\divideontimes$4)** is closely-related to the rule Cautious Monotony from the area of non-monotonic reasoning [15], while **($\divideontimes$5)** and **($\divideontimes$6)** are the two AGM supplementary postulates for contraction [1].

Note the non-appearance in this list of the AGM contraction postulates Vacuity ($\divideontimes(\bot) \nvdash \lambda$ implies $\divideontimes(\lambda) \equiv \divideontimes(\bot)$), Inclusion ($\divideontimes(\bot) \vdash \divideontimes(\lambda)$) and Recovery ($\divideontimes(\lambda) \wedge \lambda \vdash \divideontimes(\bot)$), none of which are valid in general for basic removal. Inclusion has been questioned as a general requirement for removal in [6], while Recovery has long been noted as controversial (see, e.g., [10]). Vacuity is a little harder to argue against. It says if the sentence to be removed is not in the intial belief set, then the belief set should remain unchanged. Nevertheless we feel there *are* plausible removal scenarios in which it may fail, one of which will be described in Section 3.1 below when we introduce the subclass of *prioritised* removal functions. For basic removals Inclusion actually implies Vacuity [7].

Note: The postulates are the same ones as in [7], but their appearance is changed to take into account the fact we take $\divideontimes$ to be a unary operator which returns a sentence (rather than a logically-closed set of sentences). We also leave out one rule from the list in [7], which in our reformulation corresponds to "$\divideontimes(\bot) \wedge \neg\lambda \vdash \divideontimes(\lambda)$". This rule turns out to be redundant, being derivable mainly from **($\divideontimes$3)**.

As well as the above postulates, [7] also gave a semantic account of basic removal. A *context* is any pair $\mathcal{C} = (\leq, \preceq)$ of binary relations over W such that *(i)* $\leq$ is a total preorder, i.e., transitive and connected, and *(ii)* $\preceq$ is a reflexive sub-relation of $\leq$. From any such $\mathcal{C}$ we may define a removal operator $\divideontimes_\mathcal{C}$ by setting

$$[\divideontimes_\mathcal{C}(\lambda)] = \{w \in W \mid w \preceq w' \text{ for some } w' \in \min_\leq([\neg\lambda])\}.$$

That is, the set of worlds following removal of λ is determined by first locating the $\leq$-minimal worlds in $[\neg\lambda]$, and then taking along with these all worlds which are less than them according to $\preceq$. We call $\divideontimes_{\mathcal{C}}$ the removal function *generated by* $\mathcal{C}$. [7] showed $\divideontimes_{\mathcal{C}}$ is a basic removal function and that in fact *every* basic removal function is generated from a unique context. For another, closely-related, family of belief removal functions see [8].

In this paper, another property which we will find useful, especially for technical reasons, is *Hyperregularity* [12]:

$$\text{If } \divideontimes(\lambda \wedge \chi) \not\vdash \lambda \text{ then } \divideontimes(\lambda \wedge \chi) \equiv \divideontimes(\lambda).$$

This rule says if the removal of $\lambda \wedge \chi$ excludes λ then removing $\lambda \wedge \chi$ is the same as removing just λ. This property is very strong. Not only does it imply Vacuity, but in the presence of ($\divideontimes$**1**) and ($\divideontimes$**2**) it implies ($\divideontimes$**3**)-($\divideontimes$**6**). It is probably *too* strong to be required in general. Indeed given ($\divideontimes$**1**) and ($\divideontimes$**2**) it can be shown to imply the"Decomposition" property of removal, i.e, either $\divideontimes(\lambda \wedge \chi) \equiv \divideontimes(\lambda)$ or $\divideontimes(\lambda \wedge \chi) \equiv \divideontimes(\chi)$, which has been noted as overly strong in [9, p66]. Despite this it is nevertheless still satisifed by several interesting sub-classes of basic removal (see Section 3.1 below), and when proving results we will sometimes find it a useful stepping-stone towards the more general basic removal. In terms of contexts, it corresponds to requiring the following condition on $(\leq, \preceq)$, for all $w_1, w_2, w_3 \in W$:

($\mathcal{C}$-hyp) If $w_1 \preceq w_2$ and $w_2 \sim w_3$ then $w_1 \preceq w_3$

(where $\sim$ is the symmetric closure of $\leq$), i.e, whether or not $w_1 \preceq w_2$ depends only on the $\leq$-rank of w_2.

Definition 3 *A* hyperregular removal function *is any basic removal function satisfying Hyperregularity.*

In [7] it was shown that hyperregular removal functions correspond precisely to the class of *linear liberation* operators from [6].

3.1 Some examples of basic removal functions

We now give three concrete families of operators, all of which come under the umbrella of basic removal. These families will be useful when we come to describing examples of equilibria.

(i). **Prioritised removal** Let $\langle \Sigma, \sqsubseteq \rangle$ be any finite set of *consistent* sentences Σ, totally preordered by a relation $\sqsubseteq$ over Σ. Intuitively the different sentences in Σ correspond to different possible *extensions*, prioritised by $\sqsubseteq$ (and with sentences lower down in the ordering given higher priority). Given such a set, for any $\lambda \in L_*$ let $\Sigma(\lambda) = \{\gamma \in \Sigma \mid \gamma \not\vdash \lambda\}$. Then we define $\divideontimes_{\langle \Sigma, \sqsubseteq \rangle}$ from $\langle \Sigma, \sqsubseteq \rangle$ by setting:

$$\divideontimes_{\langle \Sigma, \sqsubseteq \rangle}(\lambda) = \begin{cases} \bigvee \min_{\sqsubseteq} \Sigma(\lambda) & \text{if } \bigvee \Sigma \not\vdash \lambda \\ \top & \text{otherwise.} \end{cases}$$

In other words, after removing λ, the new belief set is just the disjunction of all the $\sqsubseteq$-minimal elements in Σ which do not entail λ. In case there is no sentence in Σ which fails to imply λ, then the result is just $\top$. We will call any removal function definable in this way a *prioritised removal* function. A similar family of removal has also been studied in [4].

One can easily check that $\divideontimes_{\langle \Sigma, \sqsubseteq \rangle}$ satisfies ($\divideontimes$**1**)-($\divideontimes$**6**) and so forms a basic removal function. Note however that $\divideontimes_{\langle \Sigma, \sqsubseteq \rangle}$ will fail to satisfy Vacuity (hence also Hyperregularity) in general. For example suppose $\Sigma = \{p, \neg p\}$ but $\sqsubseteq$ is the "flat" ordering on Σ which ranks both sentences equally. This would correspond to a situation in which an agent has equally good reasons to believe p and $\neg p$. The belief set corresponding to this is then $\divideontimes_{\langle \Sigma, \sqsubseteq \rangle}(\bot) = p \vee \neg p$, i.e., since the agent cannot choose between p and $\neg p$, he commits to neither. But $\divideontimes_{\langle \Sigma, \sqsubseteq \rangle}(p) = \neg p$. That is, the direction to remove p tips the balance in favour of $\neg p$, and the agent thus comes to believe $\neg p$, even though p was not in the initial belief set. We take this plausible removal scenario as indication that the Vacuity rule may be too strong in general.

(ii). **Severe withdrawal** [22]. A *severe withdrawal* function may be represented by a *logical chain* $\rho = \beta_1 \vdash \beta_2 \vdash \cdots \vdash \beta_m$, with $\divideontimes_\rho(\lambda) = \beta_i$, where i is minimal such that $\beta_i \nvdash \lambda$ (equals $\top$ if no such i exists). Severe withdrawal functions always satisfy Inclusion and Hyperregularity. It is easy to see they form a special case of prioritised removal. Severe withdrawal functions also have a simple representation in terms of their generating contexts $(\leq, \preceq)$. They are just those basic removals for which $\leq = \preceq$.

(iii). σ**-liberation** [6]. *σ-liberation* functions again use a sequence of sentences $\sigma = (\alpha_1, \ldots, \alpha_s)$. Given such σ and $\lambda \in L_*$, define a sequence of sentences $f_i(\sigma, \lambda)$ inductively on i by setting $f_0(\sigma, \lambda) = \top$, and then for $i > 0$,

$$f_i(\sigma, \lambda) = \begin{cases} f_{i-1}(\sigma, \lambda) \wedge \alpha_i & \text{if } f_{i-1}(\sigma, \lambda) \wedge \alpha_i \nvdash \lambda \\ f_{i-1}(\sigma, \lambda) & \text{otherwise.} \end{cases}$$

In other words, $f_s(\sigma, \alpha)$ is obtained by starting with $\top$, and then working through σ from left to right, adding each sentence provided doing so does not lead to the inference of λ. (In [6] the direction was right-to-left, but this difference is inessential.) Then $\divideontimes_\sigma(\lambda) = f_s(\sigma, \lambda)$. (This is very closely-related to the "linear base-revision" of [19].) σ-liberation functions do not satisfy Inclusion in general, but they do satisfy Hyperregularity (and hence also Vacuity). In terms of their generating contexts, σ-liberation functions correspond to those contexts $(\leq, \preceq)$ which satisfy the Hyperregularity condition **($\mathcal{C}$-hyp)** and for which $\preceq$ is transitive.

The three families described above are inter-related as follows: severe withdrawal $\subset$ σ-liberation $\subset$ prioritised removal. The inclusions are strict. In addition to these three, [7] showed basic removal includes many other well-known families of removal functions, including systematic withdrawal [17], AGM contraction and even AGM *revision*. **In the rest of the paper we shall assume the domain of a social removal function is the set of all n-tuples of basic removal functions.**

4 Removal equilibria

When is the outcome of an operation of social removal in *equilibrium*? Our first idea is the following.

Definition 4 $(\phi_i)_{i \in \mathbb{A}}$ *is a* removal equilibrium *for* $(\divideontimes_i)_{i \in \mathbb{A}}$ *iff it is consistent and, for each* $i \in \mathbb{A}$, $\phi_i \equiv \divideontimes_i(\neg \bigwedge_{j \neq i} \phi_j)$.

This definition is a direct formulation of the idea that each agent removes precisely the "right" sentence to be consistent with every other agent. As such this seems like a good candidate for a first formalisation of the word "appropriate" in our Principle of Equilibrium from the introduction.

Example 1 *Assume* $\mathbb{A} = \{1, 2\}$ *and suppose both agents use severe withdrawal to remove beliefs. Let* $\divideontimes_1$ *and* $\divideontimes_2$ *be specified by the logical chains* $(p \wedge q) \vdash q$ *and* $(\neg p \wedge \neg q) \vdash (\neg p \vee \neg q)$ *resp. Then there are three possible removal equilibria for the profile* $(\divideontimes_1, \divideontimes_2)$*: (1)* $(p \wedge q, \top)$*, corresponding to a case where 1 removes nothing and 2 removes everything, (2)* $(\top, \neg p \wedge \neg q)$*, corresponding to the opposite case, and (3)* $(q, \neg p \vee \neg q)$*, corresponding to the case where both agents give up* something*, but not everything.*

We might be interested in requiring the following property for social removal functions:

(FREq) $\mathbf{F}((\divideontimes_i)_{i \in \mathbb{A}})$ is a removal equilibrium for $(\divideontimes_i)_{i \in \mathbb{A}}$.

Is **(FREq)** even consistent? In other words, do removal equilibria always exist for *any* profile of basic removal functions? We shall shortly answer this question in the affirmative. But before that we examine such equilibria in the special case when $(\divideontimes_i(\bot))_{i \in \mathbb{A}}$ is consistent, and examine the compatibility of **(FREq)** with **(FVac)** and **(FVac$_\wedge$)**. First, the following example shows **(FREq)** is *not* compatible with **(FVac)**.

Example 2 *Again suppose* $\mathbb{A} = \{1, 2\}$*. Suppose agent 1 uses the prioritised removal function* $\divideontimes_{\langle \Sigma, \sqsubseteq \rangle}$ *where* $\Sigma = \{p, \neg p\}$ *and* $\sqsubseteq$ *is the flat priority ordering, and suppose agent 2 uses the severe withdrawal function specified by the single element logical chain* (p)*. We have* $\divideontimes_1(\bot) \equiv \top$

and $*_2(\bot) = p$. *Then* $*_1(\bot) \wedge *_2(\bot)$ *is equivalent to* p *and so is clearly consistent, but* $(*_1(\bot), *_2(\bot))$ *is not a removal equilibrium. This is because, while we do have* $*_2(\neg\top) \equiv p$, *we have* $*_1(\neg p) \equiv p \not\equiv \top$.

Thus for general basic removal profiles, we cannot require **both** (**FREq)** and (**FVac)**. At first glance it might be thought (**FVac)** is unquestionable, and so it is (**FREq)** which must be given up. However we believe that as soon as one takes the step – as we do – to relax Vacuity for *individual* removal $*$, then (**FVac)** itself becomes less "untouchable". Thus we believe this incompatibility with (**FVac)** should not by itself be taken as reason to reject (**FREq)**. Furthermore the next result (which may be proved using the same construction as in Prop. 9 below) shows (**FREq)** *is* compatible with (**FVac**$_\wedge$).

Proposition 1 *If* $(*_i(\bot))_{i\in\mathbb{A}}$ *is consistent then there exists a removal equilibrium* $(\phi_i)_{i\in\mathbb{A}}$ *for* $(*_i)_{i\in\mathbb{A}}$ *such that* $(\phi_i)_{i\in\mathbb{A}} \equiv_\wedge (*_i(\bot))_{i\in\mathbb{A}}$.

In Example 2 we do indeed have a removal equilibrium which is conjunction-equivalent to $(*_1(\bot), *_2(\bot))$, namely (p, p).

Note that in Example 2, agent 1 uses a removal function which does not satisfy Vacuity. The next result says that if we *do* insist on Vacuity for individual removal functions, then we do achieve compatibility with (**FVac)**.

Proposition 2 *Suppose each* $*_i$ *satisfies Vacuity, and suppose* $(*_i(\bot))_{i\in\mathbb{A}}$ *is consistent. Then* $(*_i(\bot))_{i\in\mathbb{A}}$ *is a removal equilibrium for* $((*_i)_{i\in\mathbb{A}})$.

However, even if the $*_i$ satisfy Vacuity, this might not be the *only* removal equilibrium. That is, even in this restricted domain case, (**FREq)** is not enough by itself to imply (**FVac)** or even (**FVac**$_\wedge$).

Example 3 *Let* $*$ *be the* σ*-liberation function determined by the sequence* $(p, \neg p)$. *Then the belief set associated to* $*$ *is* $*(\bot) = p$. *Now suppose we have* n *agents, all using this* same *removal function* $*$. *Then for the resulting removal profile there are two removal equilibria. As well as the expected* $(p)_{i\in\mathbb{A}}$ *we also get* $(\neg p)_{i\in\mathbb{A}}$*!*

It might seem bizarre that $(\neg p)_{i\in\mathbb{A}}$ should be recognised as an equilibrium in this example. Why should the agents all jump across to $\neg p$ when they can just as well stay with the comfort of p? In fact the situation is analogous to that with Nash equilibrium itself. We shall expand on this point later after we introduce the notion of entrenchment equilibria.

By restricting the domain of $\mathbf{F}$ further, we *do* force a unique removal equilibrium in the case when the initial belief sets are jointly consistent.

Proposition 3 *Suppose each* $*_i$ *satisfies Inclusion (and hence also Vacuity). Then if* $(*_i(\bot))_{i\in\mathbb{A}}$ *is consistent then it is the* only *removal equilibrium for* $(*_i)_{i\in\mathbb{A}}$.

5 Existence of removal equilibria

In this section we prove that removal equilibria are guaranteed to exist when the agents use basic removal functions to remove beliefs. First we concentrate on the case when all agents use hyperregular removal, providing two concrete social removal operators which satisfy (**FREq)**. We will build on this case to prove existence in the general basic removal case.

5.1 The hyperregular case: First method

Our first social removal function $\mathbf{F}_1$ requires the upfront specification of a linear order on $\mathbb{A}$. Without loss we take this order here to be just the numerical one on $\mathbb{A} = \{1, 2, \ldots, n\}$. Given a removal profile $(*_i)_{i\in\mathbb{A}}$, we define $\mathbf{F}_1((*_i)_{i\in\mathbb{A}}) = (\phi_i)_{i\in\mathbb{A}}$ inductively by setting

$$\phi_i = *_i(\neg \bigwedge_{j<i} \phi_j).$$

In other words, ϕ_1 is just taken to be agent 1's initial belief set $*_1(\bot)$, and then each agent takes his turn to remove the negation of the conjunction of the belief sets of all those agents whose turn has already passed. By an easy induction on i, and using the fact each $*_i$ satisfies **(*1)**, we know $\neg \bigwedge_{j<i} \phi_j \in L_*$ and so $*_i(\neg \bigwedge_{j<i} \phi_j)$ is well-defined. In particular we know from **(*1)** that $\phi_n = *_n(\neg \bigwedge_{j<n} \phi_j) \not\vdash \neg \bigwedge_{j<n} \phi_j$ and so $(\phi_i)_{i\in\mathbb{A}}$ is consistent.

Proposition 4 *If all $*_i$ satisfy Hyperregularity then $\mathbf{F}_1$ returns a removal equilibrium for $(*_i)_{i\in\mathbb{A}}$.*

$\mathbf{F}_1$ might not return a removal equilibrium for general basic removal profiles. This can be seen on Example 2, where running the above procedure returns the non-equilibrium $(\top, p)$.

What other properties does $\mathbf{F}_1$ satisfy? Well to begin, it can be shown to satisfy **(FVac)** (in the hyperregular case). Also, let's say two removal functions $*$ and $*'$ are *revision-equivalent* iff $*(\lambda) \wedge \neg\lambda \equiv *'(\lambda) \wedge \neg\lambda$ for all $\lambda \in L_*$. (i.e., the revision functions defined from them via the Levi Identity [16] are the same). Then we have:

Proposition 5 $\mathbf{F}_1$ *satisfies the following rule for social removal functions:*

*(***FRev**$_\wedge$*)* *If $*_i$ and $*'_i$ are revision-equivalent for each $i \in \mathbb{A}$ then $\mathbf{F}((*_i)_{i\in\mathbb{A}}) \equiv_\wedge \mathbf{F}((*'_i)_{i\in\mathbb{A}})$.*

In fact $\mathbf{F}_1$ satisfies this property even in the general basic removal case. Letting $\mathbf{F}_1((*_i)_{i\in\mathbb{A}}) = (\phi_i)_{i\in\mathbb{A}}$ and $\mathbf{F}_1((*'_i)_{i\in\mathbb{A}}) = (\phi'_i)_{i\in\mathbb{A}}$, the proof proceeds by induction on i that $\bigwedge_{j\leq i} \phi_j \equiv \bigwedge_{j\leq i} \phi'_j$. This result implies that if we are only interested in the result of *merging*, we could just focus on revision functions only.

One questionable property of $\mathbf{F}_1$ is we *always* get $\phi_1 = *_1(\bot)$ for any input removal profile. Thus agent 1 never leaves his initial belief set. He assumes a dictator-like role. Our second construction aims at rectifying this.

5.2 The hyperregular case: Second method

Our second construction is just like the first, except now, at the start of the process, agent 1 removes some fixed, possibly consistent, sentence χ (chosen independently of the given removal profile) rather than remove $\bot$ as before. Formally, the function $\mathbf{F}_2$ makes use of an auxilliary function $\mathbf{s}$ which takes as arguments a removal profile $(*_i)_{i\in\mathbb{A}}$ together with a sentence $\chi \in L_*$, and outputs a belief profile $(\eta_i)_{i\in\mathbb{A}}$. The η_i are defined inductively by setting $\eta_1 = *_1(\chi)$, and then for $i > 1$,

$$\eta_i = *_i(\neg \bigwedge_{j<i} \eta_j).$$

Note that if $\chi \equiv \bot$ then this is just $\mathbf{F}_1((*_i)_{i\in\mathbb{A}})$. Is this a removal equilibrium? In fact the result of this operation will be a removal equilibrium for agents $2, \ldots, n$, but not necessarily for agent 1.

Proposition 6 *Assume all $*_i$ satisfy Hyperregularity and let $\mathbf{s}(\chi \mid (*_i)_{i\in\mathbb{A}}) = (\eta_i)_{i\in\mathbb{A}}$. Then for each $i > 1$, $\eta_i \equiv *_i(\neg \bigwedge_{j\neq i} \eta_j)$, but in general $\eta_1 \not\equiv *_1(\neg \bigwedge_{j>1} \eta_j)$.*

In case $\eta_1 \not\equiv *_1(\neg \bigwedge_{j>1} \eta_j)$ we just try again with $\mathbf{s}(\chi \wedge \neg \bigwedge_{j>1} \eta_j \mid (*_i)_{i\in\mathbb{A}})$. Precisely, $\mathbf{F}_2$ is defined via the following iterative procedure:

1. Calculate $\mathbf{s}(\chi \mid (*_i)_{i\in\mathbb{A}}) = (\eta_i)_{i\in\mathbb{A}}$.
2. If $\eta_1 \equiv *_1(\neg \bigwedge_{j>1} \eta_j)$ then STOP and output $\mathbf{F}_2((*_i)_{i\in\mathbb{A}}) = (\eta_i)_{i\in\mathbb{A}}$. Otherwise set $\chi := \chi \wedge \neg \bigwedge_{j>1} \eta_j$ and go to step 1.

In case the termination condition in step 2 is not met, it can be shown $\chi \not\equiv \chi \wedge \neg \bigwedge_{j>1} \eta_j$, so we generate a stricly stronger sentence to input back into $\mathbf{s}(\cdot \mid (*_i)_{i\in\mathbb{A}})$ in step 1. Hence the process continues at most until we input $\bot$. But in this case $\mathbf{s}(\bot \mid (*_i)_{i\in\mathbb{A}}) = \mathbf{F}_1((*_i)_{i\in\mathbb{A}})$ as we have seen. Hence:

Proposition 7 *If all the $*_i$ satisfy Hyperregularity then $\mathbf{F}_2$ satisfies (***FREq***).*

For example, if we run this method on Example 3, taking $\chi = p$, we obtain the 2^{nd} equilibrium $\mathbf{F}_2((\circledast)_{i\in\mathbb{A}}) = (\neg p)_{i\in\mathbb{A}}$. Hence we see $\mathbf{F}_2$ does not validate (**FVac**$_\wedge$). It also does not satisfy (**FRev**$_\wedge$), since it can be shown the σ-liberation function from Example 3 is revision-equivalent to the severe withdrawal function $\circledast_\rho$ determined by the 1-element chain $\rho = (p)$. But if we again take $\chi = p$ then $\mathbf{F}_2((\circledast_\rho)_{i\in\mathbb{A}}) = (p)_{i\in\mathbb{A}}$.

Note although agent 1 no longer has dictator-like powers in $\mathbf{F}_2$, agent j still *dominates* all agents k for which $2 \leq j < k$, in the sense that if $\mathbf{F}_2((\circledast_i)_{i\mathbb{A}}) = (\phi_i)_{i\in\mathbb{A}}$, we *always* end up with $\phi_j = \circledast_j(\neg \bigwedge_{s<j} \phi_s)$. This means j *never* takes into account the beliefs of $k > j$ when calculating his new beliefs.

A natural question to ask is: is *every* removal equilibrium for $(\circledast_i)_{i\in\mathbb{A}}$ obtainable by the above iterative method for appropriate choices of ordering of agents and starting points χ? The next example shows the answer is generally no.

Example 4 *Suppose three agents, all using severe withdrawal functions specified respectively by the following logical chains:* $\circledast_1 : (p \leftrightarrow \neg q) \vdash (p \vee q)$, $\circledast_2$*:* $\neg q \vdash (p \vee \neg q)$, $\circledast_3$*:* $\neg p \vdash (\neg p \vee q)$. *Then the reader may check* $(\phi_1, \phi_2, \phi_3) = (p \vee q, p \vee \neg q, \neg p \vee q)$ *is a removal equilibrium (giving a merging result of* $\phi_1 \wedge \phi_2 \wedge \phi_3 \equiv p \wedge q$*). However, note this equilibrium has the special property that for each* i*, there is no* proper subset $X \subset \{j \in \mathbb{A} \mid j \neq i\}$ *such that* $\phi_i \equiv \circledast_i(\neg \bigwedge_{j\in X} \phi_j)$. *Hence this point cannot be reached using* $\mathbf{F}_2$*, since as we just remarked, there we* always *end up with* $\phi_2 \equiv \circledast_2(\neg\phi_1)$.

In the above example it could be said that at the point $(p \vee q, p \vee \neg q, \neg p \vee q)$ the three agents are all in a state of *perfect tension* with regard to one another. Each agent contributes equally to the equilibrium. We make the following definition:

Definition 5 *Let* $(\phi_i)_{i\in\mathbb{A}}$ *be a removal equilibrium for* $(\circledast_i)_{i\in\mathbb{A}}$*. Then it is a* perfect removal equilibrium *iff for each* i*, there is no* proper subset $X \subset \{j \in \mathbb{A} \mid j \neq i\}$ *such that* $\phi_i \equiv \circledast_i(\neg \bigwedge_{j\in X} \phi_j)$.

The next question is: do perfect removal equilibria always exist for any given removal profile? The answer is no, because according to the definition we may *not* have $\phi_i \equiv \circledast_i(\neg \bigwedge_{j\in\emptyset} \phi_j)$, i.e., we may not have $\phi_i \equiv \circledast_i(\bot)$. However, we may conceive of examples in which, for *every* removal equilibrium there exists at least one agent i for which $\phi_i \equiv \circledast_i(\bot)$. Indeed this will typically happen in the case of *drastic* removal profiles, see Section 7 below.

5.3 Existence: The general case

We have established that if all agents use hyperregular removal, then removal equilibria are guaranteed to exist. We now extend this fact to the case of arbitrary basic removal profiles. Given an arbitrary $(\circledast_i)_{i\in\mathbb{A}}$, we first convert each $\circledast_i$ to its *hyperregular version* $\circledast_i^h$, and then show that every removal equilibrium for $(\circledast_i^h)_{i\in\mathbb{A}}$ can be *converted* into an equilibrium for the original profile. To do this we go back to the semantic representation of basic removal functions which was mentioned after Defn. 2.

Definition 6 *Let* $\circledast$ *be a basic removal function and* $(\leq, \preceq)$ *its generating context. Then the* hyperregular version *of* $\circledast$ *is the removal operator* $\circledast^h$ *generated by* $(\leq, \preceq^h)$*, where* $\preceq^h$ *is defined by:* $w_1 \preceq^h w_2$ *iff* $w_1 \preceq w_3$ *for some* w_3 *s.t.* $w_3 \sim w_2$ *(where* $\sim$ *is the symmetric closure of* $\leq$*).*

The following are the relevant properties of $\circledast^h$:

Proposition 8 (i). $\circledast^h$ *satisfies Hyperregularity.* (ii). *For all* $\lambda \in L_*$, $\circledast(\lambda) \vdash \circledast^h(\lambda)$. (iii). $\circledast$ *and* $\circledast^h$ *are revision-equivalent.*

Now, suppose we start with arbitrary $(\circledast_i)_{i\in\mathbb{A}}$ and suppose we have found some removal equilibrium $(\phi'_i)_{i\in\mathbb{A}}$ **for the hyperregular versions** $(\circledast_i^h)_{i\in\mathbb{A}}$. Then for each i set

$$\phi_i = \circledast_i(\neg(\bigwedge_{j<i} \phi_j \wedge \bigwedge_{j>i} \phi'_j)).$$

Proposition 9 *$(\phi_i)_{i\in\mathbb{A}}$ is a removal equilibrium for $(\circledast_i)_{i\in\mathbb{A}}$. Furthermore $(\phi_i)_{i\in\mathbb{A}} \equiv_\wedge (\phi'_i)_{i\in\mathbb{A}}$.*

The second part of this proposition implies that if we are interested only in the result of *merging*, we might as well just use the Hyperregular versions.

6 Entrenchment equilibria

In this section we investigate another equilibrium notion for social belief removal, which is more directly comparable to the usual notion of Nash equilibrium in strategic games. To do so we will first show how any removal profile $(\circledast_i)_{i\in\mathbb{A}}$ defines a particular strategic game $\mathcal{G}((\circledast_i)_{i\in\mathbb{A}})$ and then use the Nash equilibria of this game to define our new notion of equilibrium. We start by recalling the definitions of strategic game and Nash equilibrium. (See, e.g., [20].)

Definition 7 *A* strategic game (over $\mathbb{A}$) *is a pair $\langle (A_i)_{i\in\mathbb{A}}, (\precsim_i)_{i\in\mathbb{A}}\rangle$, where, for each $i \in \mathbb{A}$:*

- *A_i is the set of* actions *available to agent i,*
- *$\precsim_i$ is a total preorder over $\times_{i\in\mathbb{A}} A_i$, i.e., the* preference relation *of agent i.*

The set $\times_{i\in\mathbb{A}} A_i$ is the set of *action profiles* for the agents in $\mathbb{A}$, i.e., the set of tuples consisting of a chosen action $a_i \in A_i$ for each agent i. Given two action profiles $(a_i)_{i\in\mathbb{A}}$ and $(b_i)_{i\in\mathbb{A}}$, $(a_i)_{i\in\mathbb{A}} \precsim_j (b_i)_{i\in\mathbb{A}}$ means agent j prefers (the outcome resulting from) the action profile $(b_i)_{i\in\mathbb{A}}$ *at least as much as* $(a_i)_{i\in\mathbb{A}}$.

Definition 8 *A* Nash equilibrium of a strategic game $\langle (A_i)_{i\in\mathbb{A}}, (\precsim_i)_{i\in\mathbb{A}}\rangle$ *is an action profile $(a^*_i)_{i\in\mathbb{A}}$ such that, for each $j \in \mathbb{A}$, and any $a_j \in A_j$ we have $(a_i)_{i\in\mathbb{A}} \precsim_j (a^*_i)_{i\in\mathbb{A}}$, where $a_i = a^*_i$ for $i \neq j$.*

In a Nash equilibrium no single agent can change his action in a way which leads to a more preferred outcome for him, *given* that the other agents' actions remain fixed.

How can we define a strategic game from a removal profile? Well first note in our situation of social belief removal too each agent takes an action – he chooses which sentence to remove. That is, the set of possible actions of agent i may be identified with L_*. What, then, is the preference relation of agent i over the resulting set of action profiles $\times_{j\in\mathbb{A}} L_*$? Clearly each agent prefers any action profile leading to a consistent outcome over one which leads to inconsistency. But what is his preference between different profiles leading to consistent outcomes? One natural idea is that agents prefer to remove *less entrenched* sentences [9]. Given agent i is using removal function $\circledast_i$, his *entrenchment ordering* (over L_*) $\trianglelefteq^{\mathrm{E}}_i$ is given by

$$\lambda \trianglelefteq^{\mathrm{E}}_i \chi \text{ iff } \circledast_i (\lambda \wedge \chi) \not\vdash \lambda.$$

Thus χ is *at least as entrenched as* λ iff the removal of the conjuction causes λ to be excluded. It expresses that agent i finds it *at least as easy to discard* λ as χ.

Proposition 10 *If $\circledast_i$ is a basic removal function, and $\trianglelefteq^{\mathrm{E}}_i$ is defined from $\circledast_i$ as above then $\trianglelefteq^{\mathrm{E}}_i$ forms a standard entrenchment ordering in the sense of [9]. In particular $\trianglelefteq^{\mathrm{E}}_i$ is a total preorder over L_*.*

Given this, agent i's preference relation $\precsim^{\mathrm{E}}_i$ over the set $\times_{j\in\mathbb{A}} L_*$ is may be specified completely as follows. Given any two action profiles $(\lambda_j)_{j\in\mathbb{A}}$ and $(\chi_j)_{j\in\mathbb{A}}$, we set $(\lambda_j)_{j\in\mathbb{A}} \precsim^{\mathrm{E}}_i (\chi_j)_{j\in\mathbb{A}}$ iff one of the following two conditions holds:

either *(i).* $(\circledast_j(\lambda_j))_{j\in\mathbb{A}}$ is inconsistent
or *(ii).* $(\circledast_j(\lambda_j))_{j\in\mathbb{A}}$ and $(\circledast_j(\chi_j))_{j\in\mathbb{A}}$ are both consistent and $\chi_i \trianglelefteq^{\mathrm{E}}_i \lambda_i$.

Since $\trianglelefteq^{\mathrm{E}}_i$ is a total preorder over L_*, it is easy to check $\precsim^{\mathrm{E}}_i$ forms a total preorder over the set of all action profiles.

Definition 9 *Given a removal profile $(\circledast_i)_{i\in\mathbb{A}}$, the strategic game $\langle (L_*)_{i\in\mathbb{A}}, (\precsim^{\mathrm{E}}_i)_{i\in\mathbb{A}}\rangle$ defined from $(\circledast_i)_{i\in\mathbb{A}}$ as above will be denoted by $\mathcal{G}((\circledast_i)_{i\in\mathbb{A}})$.*

Given all this, we are ready to define our next equilibrium notion.

Definition 10 $(\phi_i)_{i\in\mathbb{A}}$ *is an* entrenchment equilibrium *for* $(\divideontimes_i)_{i\in\mathbb{A}}$ *iff it is consistent and* $(\phi_i)_{i\in\mathbb{A}} \equiv (\divideontimes_i(\lambda_i^*))_{i\in\mathbb{A}}$ *for some Nash equilibrium* $(\lambda_i^*)_{i\in\mathbb{A}}$ *of the game* $\mathcal{G}((\divideontimes_i)_{i\in\mathbb{A}})$.

Put more directly, an entenchment equilibrium is an outcome $(\phi_i)_{i\in\mathbb{A}}$ which is consistent and for which no *single* agent may deviate and remove a less entrenched sentence *without* destroying this consistency. This brings us to the following social removal property:

(FEEq) $\mathbf{F}((\divideontimes_i)_{i\in\mathbb{A}})$ is an entrenchment equilibrium for $(\divideontimes_i)_{i\in\mathbb{A}}$.

What is the relationship between entrenchment equilibria and removal equilibria?

Proposition 11 *Every removal equilibrium for* $(\divideontimes_i)_{i\in\mathbb{A}}$ *is an entrenchment equilibrium for* $(\divideontimes_i)_{i\in\mathbb{A}}$. *Furthermore if all* $\divideontimes_i$ *are hyperregular then every entrenchment equilibrium for* $(\divideontimes_i)_{i\in\mathbb{A}}$ *is a removal equilibrium for* $(\divideontimes_i)_{i\in\mathbb{A}}$.

Thus if all agents use hyperregular removal then the two notions of equilibrium *coincide*. However, in general, not every entrenchment equilibrium is a removal equilibrium, since for example if $(\divideontimes_i(\bot))_{i\in\mathbb{A}}$ is consistent then it is *always* an entrenchment equilibrium, because $\bot$ is always minimally entrenched for any basic removal function. However we have already seen that it might not be a removal equilibrium.

As we saw in Example 3, even in the hyperregular case, if $(\divideontimes_i(\bot))_{i\in\mathbb{A}}$ is consistent it might still not be the *only* entrenchment equilibrium. It might seem irrational for both agents to give up p in this example, when it's possible for both to remove a less entrenched sentence (i.e. $\bot$) while preserving consistency. This kind of counterintuitive result is not restricted to entrenchment equilibria. In fact it is inherent in the concept of Nash equilibrium itself. It has long been recognised that the Nash equilibrium does not rule out sub-optimal solutions in the case where agents have identical preferences over outcomes. This is illustrated by the following example, taken from [20, p16].

Example 5 *Suppose two agents* $\{1, 2\}$ *who wish to go to a concert together, but must choose between going to a Mozart (Mo) concert or a Mahler (Ma) concert. Thus the set of actions for both agents is* $A = \{Mo, Ma\}$. *We assume both agents have identical preferences over the four possible action profiles. Firstly, the agents want to reach agreement, so the two profiles in which they choose different actions are the least preferred. Moreover, both agents prefer to see the Mozart concert. Thus the preference relation* $\precsim$ *of both agents is specified completely by*

$$(Mo, Ma) \sim (Ma, Mo) \prec (Ma, Ma) \prec (Mo, Mo).$$

(Just for this example we are using $\sim$ *and* $\prec$ *to denote the symmetric closure and strict part of* $\precsim$ *respectively.) In this game there are two Nash equilibria* (Ma, Ma) *and* (Mo, Mo). *Even though both agents have a mutual interest in reaching* (Mo, Mo), *the Nash equilibrium does not rule out the inferior outcome* (Ma, Ma).

This anomaly led several authors to propose refined equilibria concepts for strategic games. One such refinement, the *strong* Nash equilibrium [3], says roughly that no *set* – not just singletons as with Nash – of agents can make a joint change in strategy which leads to a more preferred outcome for all agents in that set.

Definition 11 *A* strong Nash equilibrium of a strategic game $\langle (A_i)_{i\in\mathbb{A}}, (\precsim_i)_{i\in\mathbb{A}}\rangle$ *is an action profile* $(a_i^*)_{i\in\mathbb{A}}$ *such that, for any* $X \subseteq \mathbb{A}$, *and each tuple* $(a_i)_{i\in X}$, *there exists* $j \in X$ *such that* $(a_i)_{i\in\mathbb{A}} \precsim_j (a_i^*)_{i\in\mathbb{A}}$, *where* $a_i = a_i^*$ *for* $i \notin X$.

This leads to the corresponding refinement for entrenchment equilibria.

Definition 12 $(\phi_i)_{i\in\mathbb{A}}$ *is a* strong entrenchment equilibrium *for* $(\divideontimes_i)_{i\in\mathbb{A}}$ *iff it is consistent and* $(\phi_i)_{i\in\mathbb{A}} \equiv (\divideontimes_i(\lambda_i^*))_{i\in\mathbb{A}}$ *for some strong Nash equilibrium* $(\lambda_i^*)_{i\in\mathbb{A}}$ *of the game* $\mathcal{G}((\divideontimes_i)_{i\in\mathbb{A}})$.

The following property thus strengthens (**FEEq**):

(**FEEq+**) $\mathbf{F}((\divideontimes_i)_{i\in\mathbb{A}})$ is a strong entrenchment equilibrium for $(\divideontimes_i)_{i\in\mathbb{A}}$.

In Example 3 the only strong entrenchment equilibrium is $(p)_{i\in\mathbb{A}}$. For hyperregular removal profiles, it can be shown function $\mathbf{F}_1$ defined earlier satisfies (**FEEq+**), but $\mathbf{F}_2$ does not. Thus strong entrenchment equilibria *always* exist for hyperregular removal profiles. However at the time of writing it is an open problem whether they always exist for general basic removal profiles. It would also be interesting to try and find a necessary and sufficient condition for a removal equilibrium to be a strong entrenchment equilibrium (even in the hyperregular case).

7 Equilibria as maxiconsistent sets

The simplest kind of removal function is what might be termed *drastic removal*, in which the result of removing λ is $\divideontimes(\bot)$ if λ is not entailed by the initial belief set, or $\top$ if it *is* entailed. That is, an agent either leaves his belief set unchanged, or throws out *all* beliefs. Drastic removals correspond to the severe withdrawal functions determined by single-element logical chains.

If all agents use drastic removal, then removal/entrenchment equilibria reduce to taking *maximal consistent sets of agents*. $X \subseteq \mathbb{A}$ is maximally consistent iff *(i)* $\bigwedge_{i\in X} \divideontimes_i(\bot)$ is consistent, and *(ii)* $\bigwedge_{i\in Y} \divideontimes_i(\bot)$ is inconsistent for all $X \subset Y \subseteq \mathbb{A}$.

Proposition 12 *Suppose all $\divideontimes_i$ are drastic removal functions. Then $(\phi_i)_{i\in\mathbb{A}}$ is a removal (or entrenchment) equilibrium for $(\divideontimes_i)_{i\in\mathbb{A}}$ iff $\{i \mid \phi_i \equiv \divideontimes_i(\bot)\}$ is a maximally consistent subset of $\mathbb{A}$.*

Thus we see that the main notions of equilibria studied in this paper (removal and entrenchment) can be seen as *generalisations* of the idea of taking maximal consistent sets.

8 Conclusion

We have defined several notions of equilibrium in the framework of social removal functions, formulated purely in the language of belief removal operators. Assuming all agents use basic removal functions to remove their own beliefs, we proved our equilibria are always guaranteed to exist. We gave several examples to illustrate these notions, and we showed that they generalise in some sense the idea of resolving inconsistency by taking maximal consistent subsets of agents.

For future work, we want to generalise our results to handle social removal under *integrity constraints* [14]. An *IC social removal function* is a function taking as arguments a removal profile and a consistent sentence Ψ, which returns a belief profile which is consistent *with* Ψ. The equilibrium notions described in this paper should extend to this setting. For example an IC removal equilibrium could be defined to be any belief profile $(\phi_i)_{i\in\mathbb{A}}$ for which $\phi_i \equiv \divideontimes_i(\neg(\Psi \wedge \bigwedge_{j\neq i} \phi_j))$ for all i.

Social belief removal functions have obvious similarities to *social choice rules* [2]. A social choice rule takes as input a profile of total preorders over the set of alternatives together with a given subset A of the alternatives, and outputs a subset of A – the *chosen* elements of A for the group. By conjoining the elements of the output of a social belief removal function we obtain an output of the same type as with social choice rules, but the input of a social belief removal function can be viewed as *richer* than that for social choice, since a basic removal function corresponds to a total preorder $\leq$ *plus* a reflexive sub-relation $\preceq$. It would be interesting to explore any (im)possibility theorems for social removal functions.

Acknowledgements

Thanks are due to Alexander Nittka and to several anonymous reviewers whose comments greatly helped to improve the paper. This material is based upon work supported by the National Research Foundation under Grant number 65152.

References

[1] C. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

[2] K. Arrow, A. Sen, and K. Suzumura, editors. *Handbook of Social Choice and Welfare*. Elsevier, 2002.

[3] R. Aumann. Acceptable points in general cooperative n-person games. In *Contributions to the Theory of Games, Vol. IV*, pages 287–324, 1959.

[4] A. Bochman. *A Logical Theory of Nonmonotonic Inference and Belief Change*. Springer, 2001.

[5] R. Booth. Social contraction and belief negotiation. *Information Fusion*, 7(1):19–34, 2006.

[6] R. Booth, S. Chopra, A. Ghose, and T. Meyer. Belief liberation (and retraction). *Studia Logica*, 79(1):47–72, 2005.

[7] R. Booth, S. Chopra, T. Meyer, and A. Ghose. A unifying semantics for belief change. In *Proceedings of ECAI'04*, pages 793–797, 2004.

[8] J. Cantwell. Eligible contraction. *Studia Logica*, 73:167–182, 2003.

[9] P. Gärdenfors. *Knowledge in Flux*. MIT Press, 1988.

[10] S. O. Hansson. Belief contraction without recovery. *Studia Logica*, 50(2):251–260, 1991.

[11] S. O. Hansson. Changes on disjunctively closed bases. *Journal of Logic, Language and Information*, 2:255–284, 1993.

[12] S. O. Hansson. Theory contraction and base contraction unified. *Journal of Symbolic Logic*, 58:602–625, 1993.

[13] S. Konieczny and E. Gregoire. Logic-based approaches to information fusion. *Information Fusion*, 7(1):4–18, 2006.

[14] S. Konieczny and R. Pino Pérez. Merging information under constraints: A logical framework. *Journal of Logic and Computation*, 12(5):773–808, 2002.

[15] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1991.

[16] I. Levi. *The Fixation of Belief and Its Undoing*. Cambridge University Press, Cambridge, 1991.

[17] T. Meyer, J. Heidema, W. Labuschagne, and L. Leenen. Systematic withdrawal. *Journal of Philosophical Logic*, 31(5):415–443, 2002.

[18] J. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950.

[19] B. Nebel. Base revision operations and schemes: Semantics, representation and complexity. In *Proceedings of ECAI'94*, pages 342–345, 1994.

[20] M. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, 1994.

[21] H. Rott. Preferential belief change using generalized epistemic entrenchment. *Journal of Logic, Language and Information*, 1:45–78, 1992.

[22] H. Rott and M. Pagnucco. Severe withdrawal (and recovery). *Journal of Philosophical Logic*, 28:501–547, 1999.

Richard Booth
Mahasarakham University
Faculty of Informatics
Mahasarakham 44150, Thailand
Email: richard.b@msu.ac.th

Thomas Meyer
Meraka Institute
CSIR, Pretoria
South Africa
Email: tommie.meyer@meraka.org.za

A Computational Analysis of the Tournament Equilibrium Set*

Felix Brandt, Felix Fischer, Paul Harrenstein, and Maximilian Mair

Abstract

A recurring theme in the mathematical social sciences is how to select the "most desirable" elements given a binary dominance relation on a set of alternatives. Schwartz's *tournament equilibrium set* (TEQ) ranks among the most intriguing, but also among the most enigmatic, tournament solutions proposed so far in this context. Due to its unwieldy recursive definition, little is known about TEQ. In particular, its monotonicity remains an open problem to date. Yet, if TEQ were to satisfy monotonicity, it would be a very attractive solution concept refining both the Banks set and Dutta's minimal covering set. We show that the problem of deciding whether a given alternative is contained in TEQ is NP-hard. Furthermore, we propose a heuristic that significantly outperforms the naive algorithm for computing TEQ. Early experimental results support the conjecture that TEQ is indeed monotonic.

1 Introduction

A recurring theme in the mathematical social sciences is how to select the "most desirable" elements given a binary dominance relation on a set of alternatives. Examples are diverse and include selecting socially preferred candidates in social choice settings (*e.g.*, Fishburn, 1977; Laslier, 1997), finding valid arguments in argumentation theory (*e.g.*, Dung, 1995; Dunne, 2007), determining the winners of a sports tournament (*e.g.*, Dutta and Laslier, 1999), making decisions based on multiple criteria (*e.g.*, Bouyssou et al., 2006), choosing the optimal strategy in a symmetric two-player zero-sum game (*e.g.*, Duggan and Le Breton, 1996), and singling out acceptable payoff profiles in cooperative game theory (Gillies, 1959; Brandt and Harrenstein, 2008). In social choice theory, where dominance-based solutions are most prevalent, the dominance relation can simply be defined as the pairwise majority relation, *i.e.*, an alternative a is said to dominate another alternative b if the number of individuals preferring a to b exceeds the number of individuals preferring b to a. As is well known from Condorcet's paradox (de Condorcet, 1785), the dominance relation may contain cycles and thus need not have a maximum, even if each of the underlying individual preferences does. As a consequence, the concept of maximality is rendered untenable in most cases, and a variety of so-called *solution concepts* that take over the role of maximality in non-transitive relations have been suggested (see, *e.g.*, Laslier, 1997).

The *tournament equilibrium set* (TEQ) introduced by Schwartz (1990) ranks among the most intriguing, but also among the most enigmatic, solution concepts that has been proposed for tournaments, *i.e.*, asymmetric and complete dominance relations. Due to its unwieldy recursive definition, however, preciously little is known about TEQ (Dutta, 1990; Laffond et al., 1993). In particular, whether TEQ satisfies the important property of monotonicity remains an open question to date. If it does, TEQ constitutes a most attractive tournament solution, refining both the minimal covering set and the Banks set (Laslier, 1997; Laffond et al., 1993).

Recent work in computer science has addressed the computational complexity of almost

*An earlier version of this paper appeared in the proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI).

all common solution concepts (see, *e.g.*, Woeginger, 2003; Alon, 2006; Conitzer, 2006; Brandt et al., 2007). The minimal covering set and the tournament equilibrium set, however, have remained notable exceptions. Laslier writes that "Unfortunately, no algorithm has yet been published for finding the minimal covering set or the tournament equilibrium set of large tournaments. For tournaments of order 10 or more, it is almost impossible to find (in the general case) these sets at hand" (Laslier, 1997, p.8). The minimal covering set has recently been shown to be computable in polynomial time (Brandt and Fischer, 2008). In this paper we prove that the same is not true for TEQ, unless P equals NP. We first give an arguably simpler alternative to Woeginger's (2003) NP-hardness proof for membership in the Banks set. Then the construction used in that proof is modified so as to obtain the analogous result for TEQ. In contrast to the Banks set, there is no obvious reason to suppose that the TEQ membership problem is in NP; it may very well be even harder. In the second part of the paper, we propose and evaluate a heuristic for computing TEQ that performs reasonably well on tournaments with up to 150 alternatives. Experiments further support the conjecture that TEQ is indeed monotonic.

2 Preliminaries

A *tournament* T is a pair $(A, \succ)$, where A is a finite set of *alternatives* and $\succ$ an irreflexive, anti-symmetric, and complete binary relation on A, also referred to as the *dominance relation*. Intuitively, $a \succ b$ signifies that alternative a beats b in a pairwise comparison. We write $\mathcal{T}$ for the class of all tournaments and have $\mathcal{T}(A)$ denote the set of all tournaments on a fixed set A of alternatives. If T is a tournament on A, then every subset X of A induces a tournament $T|_X = (X, \succ|_X)$, where $\succ|_X = \{(x, y) \in X \times X \colon x \succ y\}$.

As the dominance relation is not assumed to be transitive in general, there need not be a so-called *Condorcet winner*, *i.e.*, an alternative that dominates all other alternatives. A *tournament solution* S is defined as a function that associates with each tournament T on A a subset $S(T)$ of A. The definition of a tournament solution commonly includes the requirement that $S(T)$ be non-empty if T is defined on a non-empty set of alternatives and that it select the Condorcet winner if there is one (Laslier, 1997, p.37). For X a subset of A, we also write $S(X)$ for the more cumbersome $S(T|_X)$, provided that the tournament T is known from the context. A tournament solution S is said to be *monotonic* if for any two tournaments $T, T' \in \mathcal{T}(A)$ which only differ in that $b \succ a$ in T and $a \succ b$ in T', $a \in S(T)$ implies that also $a \in S(T')$, *i.e.*, reinforcing an alternative cannot cause it to be excluded from the solution set. Monotonicity is a vital property that all reasonable tournament solutions satisfy. In this paper, we will be concerned with two particular tournament solutions, the Banks set and Schwartz's tournament equilibrium set (TEQ). For a proper formal definition, however, we need some auxiliary notions and notations.

Let R be a binary relation on a set A. We write R^* for the transitive reflexive closure of R. By the *top cycle* $TC_A(R)$ we understand the maximal elements of the asymmetric part of R^*. A subset X of A is said to be *transitive* if R is transitive on X. For $X \subseteq Y \subseteq A$, X is called *maximal transitive in* Y if X is transitive and no proper superset of X in Y is. Clearly, since A is finite, every transitive set is contained in a maximal transitive set. Given a set $Z = \{Z_i\}_{i \in I}$ of pairwise disjoint subsets of A, a subset X of A will be called a *choice set for* Z if it contains precisely one element from each subset $Z_i \in Z$.

In tournaments, maximal transitive sets are also referred to as Banks trajectories. The *Banks set* $BA(T)$ of a tournament T then collects the maximal elements of the Banks trajectories.

Definition 1 (Banks set) *Let T be a tournament on A. An alternative $a \in A$ is in the* Banks set $BA(T)$ *of T if a is a maximal element of some maximal transitive set in T.*

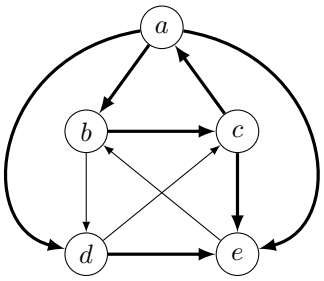


Figure 1: Example due to Schwartz, 1990, where $BA(T) = \{a, b, c, d\}$ and $TEQ(T) = \{a, b, c\}$. The TEQ relation $\longrightarrow$ is indicated by thick edges.

The *tournament equilibrium set* $TEQ(T)$ of a tournament T on A is defined as the top cycle of a particular subrelation of the dominance relation, referred to as the TEQ relation in the following. The underlying idea is that an alternative is only "properly" dominated, *i.e.*, dominated according to the subrelation, if it is dominated by an element that is selected by some tournament solution concept S. To make this idea precise, for $X \subseteq A$, we write $\overline{D}_X(a) = \{\, b \in X \colon b \succ a \,\}$ for the *dominators* of a in X, omitting the subscript when $X = A$. Thus, for each alternative a one examines the set $\overline{D}(a)$ of its dominators, and solves the subtournament $T|_{\overline{D}(a)}$ by means of the solution S. In the subrelation a is then only dominated by the alternatives in $S(\overline{D}(a))$. This of course, still leaves open the question as to the choice of the solution concept S. Now, in the case of TEQ, S is taken to be TEQ itself! This recursion is well-defined because for any $X \subseteq A$ and $a \in X$, the set $\overline{D}_X(a)$ is a *proper* subset of X. Thus, in order to determine the TEQ relation in a subtournament T, one has to calculate the TEQ of smaller and smaller subtournaments of T.

Definition 2 (Tournament equilibrium set) *Let $T \in \mathcal{T}(A)$. For each subset $X \subseteq A$, define the* tournament equilibrium set *$TEQ(X)$ for X as*

$$TEQ(X) = TC_X(\longrightarrow_X),$$

where $\longrightarrow_X$ is defined as the binary relation on X such that for all $x, y \in X$,

$$x \longrightarrow_X y \text{ if and only if } x \in TEQ(\overline{D}_X(y)).$$

Recall that in particular, $TEQ(\emptyset) = \emptyset$. The TEQ relation $\longrightarrow_X$ is a subset of the dominance relation $\succ$, and if $\overline{D}_X(x) \neq \emptyset$, then there is some $y \in \overline{D}_X(x)$ with $y \longrightarrow_X x$. Furthermore, Definition 2 directly yields a recursive algorithm to compute TEQ. Some reflection reveals that this *naive algorithm* requires time exponential in $|A|$ in the worst case.

It can easily be established that the Banks set and TEQ both select the Condorcet winner of a tournament if there is one. Moreover, in a cyclic tournament on three alternatives, the Banks set and TEQ both consist of all alternatives. In more complex tournaments, however, the Banks set and TEQ may differ. Consider, for example, the tournament T depicted in Figure 1. We first calculate the TEQ relation $\longrightarrow$. Thus, *e.g.*, for alternative e we find $\overline{D}(e) = \{a, c, d\}$, which constitutes a three-cycle, and so $TEQ(\overline{D}(e)) = \{a, c, d\}$. Accordingly, $a \longrightarrow e$, $c \longrightarrow e$, as well as $d \longrightarrow e$. Doing this for all alternatives, we find $TEQ(T) = \{a, b, c\}$ as the top cycle $TC(\longrightarrow)$ of the relation $\longrightarrow$. By contrast, the Banks set consists of the four elements a, b, c and d. *E.g.*, $d \in BA(T)$, because $\{d, c, e\}$ is a maximal transitive set with maximal element d. Nevertheless, TEQ is always included in the Banks set.

Proposition 1 (Schwartz, 1990) *Let $T = (A, \succ)$ be a tournament. Then, $TEQ(T) \subseteq BA(T)$.*

Proof: We prove by structural induction on X that $TEQ(X) \subseteq BA(X)$ for all subsets X of A. The case $X = \emptyset$ is trivial, as then $TEQ(X) = BA(X) = \emptyset$. So, assume that $TEQ(X') \subseteq BA(X')$, for all $X' \subsetneq X$. We prove that $TEQ(X) \subseteq BA(X)$ as well. To this end, consider an arbitrary $a \in TEQ(X)$. Either $\overline{D}_X(a) = \emptyset$ or $\overline{D}_X(a) \neq \emptyset$. In the former case, a is the Condorcet winner in X and therefore $a \in BA(X)$. In the latter case, $x \rightarrow_X a$ for some $x \in X$. Having assumed that $a \in TEQ(X)$, *i.e.*, $a \in TC(\rightarrow_X)$, there is also an $x' \in X$ with $a \rightarrow_X x'$. Accordingly, $a \in TEQ(\overline{D}_X(x'))$. By the induction hypothesis, also $a \in BA(\overline{D}_X(x'))$. Therefore, there is some maximal transitive set Y in $\overline{D}_X(x')$ of which a is the maximal element. Then, $Y \cup \{x'\}$ is a transitive set in X. Now let $Y' \subseteq X$ be a maximal transitive set in X containing $Y \cup \{x'\}$ with a' as maximal element. Observe that $a' \in BA(X)$. Then, $a' \succ x'$ and so $a' \in \overline{D}_X(x')$. Now consider $Y' \cap \overline{D}_X(x')$. Clearly, $Y \cap \overline{D}_X(x')$ is a transitive set in $\overline{D}_X(x')$ which contains a' as its maximal element. Moreover, $Y \subseteq Y' \cap \overline{D}_X(x')$. By maximality of Y it then follows that $Y = Y' \cap \overline{D}_X(x')$ and that $a = a'$. We may conclude that $a \in BA(X)$. □

Otherwise, little is known and much surmised about the theoretical properties of TEQ. For example, Schwartz (1990) conjectured that the top cycle of the TEQ relation is always weakly connected, a property of TEQ we will refer to as *CTC* for *connected top cycle*. Laffond et al. (1993) showed that TEQ satisfying CTC is equivalent to it having a number of useful properties. In particular, TEQ is monotonic if and only if CTC holds. Moreover, CTC implies the inclusion of TEQ in the minimal covering set (see, *e.g.*, Laslier, 1997), another appealing tournament solution. Thus, if TEQ satisfies CTC it might be considered a very strong solution concept. Otherwise, TEQ lacks the vital property of monotonicity and as such it would be severely flawed as a tournament solution.

3 An Alternative NP-Hardness Proof for Membership in the Banks Set

We begin our investigation of the computational complexity of the TEQ membership problem by giving an alternative proof for NP-hardness of the analogous problem for the Banks set. The latter was first demonstrated by Woeginger (2003) using a reduction from graph three-colorability. Our proof works by a reduction from *3SAT*, the NP-complete satisfiability problem for Boolean formulas in conjunctive normal form with exactly three literals per clause (see, *e.g.*, Papadimitriou, 1994). It is arguably simpler than Woeginger's, and a much similar construction will be used in the next section to prove NP-hardness of membership in TEQ. The tournaments used in both reductions will be taken from a special class $\mathcal{T}^*$.

Definition 3 (The class $\mathcal{T}^*$) *A tournament $(A, \succ)$ is in the class $\mathcal{T}^*$ if it satisfies the following properties. There is some odd integer $n \geq 1$, the number of* layers *in the tournament, such that $A = C \cup U_1 \cup \cdots \cup U_n$, where $C, U_1, \ldots, U_n$ are pairwise disjoint and $C = \{c_0, \ldots, c_n\}$. Each U_i is a singleton if i is even, and $U_i = \{u_i^1, u_i^2, u_i^3\}$ if i is odd. The complete and asymmetric dominance relation $\succ$ is such that for all $c_i \in C_i$, $c_j \in C_j$, $u_i \in U_i$, $u_j \in U_j$ $(0 \leq i, j \leq n)$:*

(i) $c_i \succ c_j$, *if* $i > j$,

(ii) $u_i \succ c_j$, *if* $i = j$,

(iii) $c_j \succ u_i$, *if* $i \neq j$,

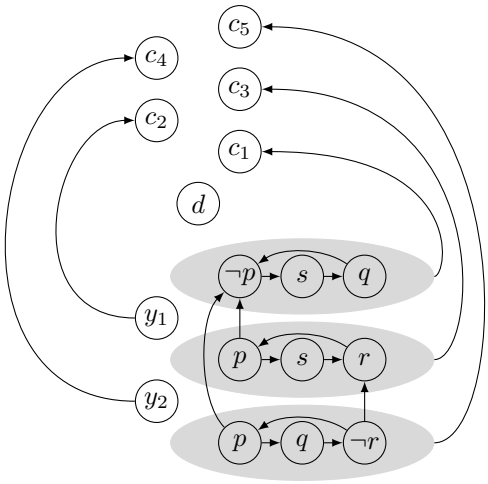


Figure 2: Tournament T_φ^{BA} for the *3CNF* formula $\varphi = (\neg p \vee s \vee q) \wedge (p \vee s \vee r) \wedge (p \vee q \vee \neg r)$. Omitted edges are assumed to point downwards.

(iv) $u_i \succ u_j$, *if* $i < j$ *and at least one of* i *and* j *is even,*

(v) $u_i^k \succ u_i^l$, *if* i *is odd and* $k \equiv l - 1 \pmod 3$

We also refer to c_0 *by* d*, for "decision node" and to* $\bigcup_{1 \leq i \leq n} U_n$ *by* U*. For* $i = 2k$*, we have as a notational convention* $U_i = Y_k = \{y_k\}$ *and set* $Y = \bigcup_{1 \leq 2k \leq n} Y_k$*.*

Observe that this definition fixes the dominance relation between any two alternatives except for some pairs of alternatives that are both in U.

As a next step in the argument, we associate with each instance of *3SAT* a tournament in the class $\mathcal{T}^*$. An instance of *3SAT* is given by a formula φ in 3-*conjunctive normal form (3CNF), i.e.*, $\varphi = (x_1^1 \vee x_1^2 \vee x_1^3) \wedge \cdots \wedge (x_m^1 \vee x_m^2 \vee x_m^3)$, where each $x \in \{x_i^1, x_i^2, x_i^3 \colon 1 \leq i \leq m\}$ is a literal. For each clause $x_i^1 \vee x_i^2 \vee x_i^3$ we assume x_i^1, x_i^2 and x_i^3 to be distinct literals. We moreover assume the literals to be indexed and by X_i we denote the set $\{x_i^1, x_i^2, x_i^3\}$. For literals x we have $\bar{x} = \neg p$ if $x = p$, and $\bar{x} = p$ if $x = \neg p$, where p is some propositional variable. We may also assume that if x and y are literals in the same clause, then $x \neq \bar{y}$. We say a *3CNF* $\varphi = (x_1^1 \vee x_1^2 \vee x_1^3) \wedge \cdots \wedge (x_m^1 \vee x_m^2 \vee x_m^3)$ is *satisfiable* if there is a choice set V for $\{X_i\}_{1 \leq i \leq m}$ such that $v' = \bar{v}$ for no $v, v' \in V$. Next we define for each *3SAT* formula φ the tournament T_φ^{BA}.

Definition 4 (Banks construction) *Let* φ *be a 3CNF* $(x_1^1 \vee x_1^2 \vee x_1^3) \wedge \cdots \wedge (x_m^1 \vee x_m^2 \vee x_m^3)$*. Define* $T_\varphi^{BA} = (C \cup U, \succ)$ *as the tournament in the class* $\mathcal{T}^*$ *with* $2m - 1$ *layers such that for all* $1 \leq j < 2m$*,*

$$U_j = \begin{cases} X_i & \text{if } j = 2i - 1, \\ \{y_i\} & \text{if } j = 2i \end{cases}$$

and such that for all $x \in X_i$ *and* $x' \in X_j$ *(*$1 \leq i, j \leq m$*),*

$$x \succ x' \quad \text{if both } j < i \text{ and } x' = \bar{x} \text{ or both } i < j \text{ and } x' \neq \bar{x}.$$

Observe that in conjunction with the other requirements on the dominance relation of a tournament in $\mathcal{T}^*$*, this completely fixes the dominance relation* $\succ$ *of* T_φ^{BA}*.*

An example of a tournament T_φ^{BA} for a *3CNF* φ is shown in Figure 2. We are now in a position to present our alternative proof that the Banks membership problem is NP-complete.

Theorem 1 *The problem of deciding whether a particular alternative is in the Banks set of a tournament is NP-complete.*

Proof: Membership in NP is obvious. For a fixed alternative d, we can simply guess a transitive subset of alternatives V with d as maximal element and verify that V is also maximal with respect to set inclusion.

For NP-hardness, we show that T_φ^{BA} contains a maximal transitive set with maximal element d if and only if φ is satisfiable. First observe that V is a maximal transitive subset with maximal element d in T_φ^{BA} only if both

(i) for all $1 \leq i < 2m$ there is a $u \in U_i$ such that $u \in V$, and

(ii) there are no $1 \leq i < j < 2m$, $u \in U_i$, $u' \in U_j$ with $u, u' \in V$ such that $u_j \succ u_i$.

Regarding (i), if there is an $1 \leq i < 2m$ such that no element of U_i is contained in V, we can always add c_i to V in order to obtain a larger transitive set. If (ii) were not to hold, both i and j have to be odd for u_j to dominate u_i. However, in light of (i), there has to be k with $i < k < j$ and $u'' \in U_k$ such that $u'' \in V$. It follows that V is not transitive because u, u'', and u' form a cycle. If there is maximal transitive set V with maximal element d complying with both (i) and (ii), a satisfying assignment of φ can be obtained by letting all literals contained in $X \cap V$ be true.

For the opposite direction, assume that φ is satisfiable. Then there is a choice set W for $\{X_i\}_{1 \leq i \leq m}$ such that $x' = \bar{x}$ for no $x, x' \in W$. Obviously $V = W \cup \{y_1, \ldots, y_{m-1}\} \cup \{d\}$ does not contain any cycles and thus is transitive with maximal element d. In order to obtain a larger transitive set with a different maximal element, we need to add c_i for some $1 \leq i \leq m$ to V. However, $V \cup \{c_i\}$ always contains a cycle consisting of c_i, d, and u for some $u \in U_i$, contradicting the transitivity of $V \cup \{c_i\}$. We have thus shown that d is the maximal element of some maximal transitive set in T_φ^{BA} containing V as a subset. □

4 NP-hardness of Membership in TEQ

In this section we prove that the problem of deciding whether a particular alternative is in the TEQ of a tournament is NP-hard. To this end, we refine the construction that was used in the previous section to prove NP-completeness of membership in the Banks set.

Definition 5 (TEQ construction) *Let φ be a 3CNF $(x_1^1 \vee x_1^2 \vee x_1^3) \wedge \cdots \wedge (x_m^1 \vee x_m^2 \vee x_m^3)$. Further for each $1 \leq i < m$, let there be a set $Z_i = \{z_i^1, z_i^2, z_i^3\}$. Define T_φ^{TEQ} as the tournament $(A, \succ)$ in $\mathcal{T}^*$ with $4n-3$ layers such that $A = C \cup U_1 \cup \cdots \cup U_{4n-3}$ and for all $1 \leq i \leq m$,*

$$U_j = \begin{cases} X_i & \text{if } j = 4i-3, \\ Z_i & \text{if } j = 4i-1, \\ \{y_i\} & \text{otherwise.} \end{cases}$$

As in the Banks construction, we let for all $x \in X_i$ and $x' \in X_j$ $(1 \leq i, j \leq m)$

$$x \succ x' \quad \text{if both } j < i \text{ and } x' = \bar{x} \text{ or both } i < j \text{ and } x' \neq \bar{x}.$$

Finally, for all $1 \leq i, j \leq m$, $x_i^k \in X_i$ and $z_j^l \in Z_j$,

$$x_i^k \succ z_j^l \text{ if and only if } i < j \text{ or both } i = j \text{ and } k = l.$$

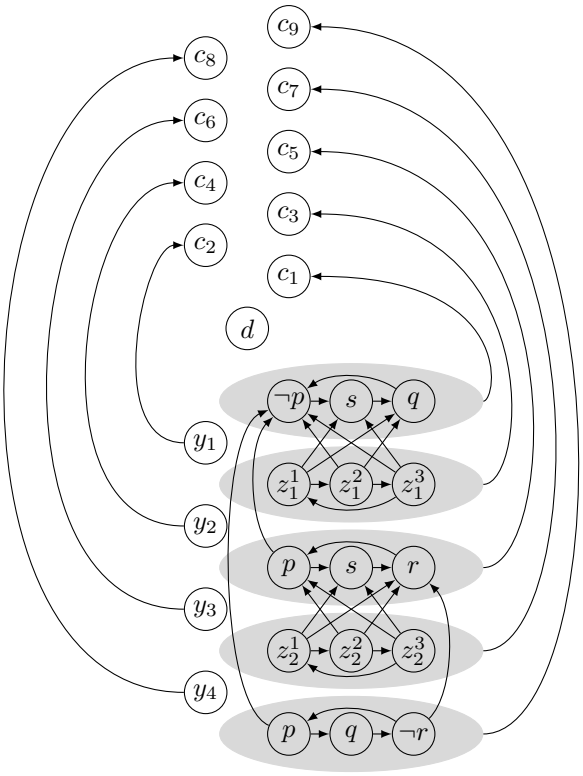


Figure 3: Tournament T_φ^{TEQ} for the *3CNF* formula $\varphi = (\neg p \vee s \vee q) \wedge (p \vee s \vee r) \wedge (p \vee q \vee \neg r)$.

An example for such a tournament is shown in Figure 3.

We now proceed to show that a *3SAT* formula φ is satisfiable if and only if the decision node d is in the tournament equilibrium set of T_φ^{TEQ}. We make use of the following lemma.

Lemma 1 *Let $T = (C \cup U, \succ)$ be a tournament in $\mathcal{T}^*$ and let $B \subseteq C \cup U$ such that $d \in B$. Then, for each $u \in U \cap B$ there exists some $c \in C \cap B$ such that $c \rightarrow_B^* u$.*

Proof: Let $c_i \in C \cap B$ be such that $\overline{D}_B(c_i) \cap C = \emptyset$, *i.e.*, c_i is the alternative in C with the highest index among those included in B. Then,

$$c_i \rightarrow_B c \text{ for all } c \in B \cap C \text{ with } c \neq c_i. \tag{1}$$

For this, merely observe that by construction c_i is the Condorcet winner in $\overline{D}_B(c)$. Hence, $c_i \in TEQ(\overline{D}_B(c))$ and $c_i \rightarrow_B c$.

The lemma itself then follows from the stronger claim that for each $u \in U \cap B$ there is some $c \in C \cap B$ with both $c \rightarrow_B^* u$ and $c \in TEQ(B)$. This claim we prove by structural induction on supersets B of $\{d\}$.

If $B = \{d\}$, $U \cap B = \emptyset$ and the claim is satisfied trivially. So let $\{d\}$ be a proper subset of B. Again, if $U \cap B = \emptyset$, the claim holds trivially. So we may assume there be some $u \in U \cap B$. Then, $d \in \overline{D}_B(u)$ by construction of T. If $\overline{D}_B(u) \cap U = \emptyset$, $\overline{D}_B(u)$ is a non-empty subset of $C \cap B$, and so is $TEQ(\overline{D}_B(u))$. It follows that for some $c \in TEQ(\overline{D}_B(u)) \cap C$ we have $c \rightarrow_B u$. If, on the other hand, $\overline{D}_B(u) \cap U \neq \emptyset$, the induction hypothesis is applicable and we have $c \in TEQ(\overline{D}_B(u))$ for some $c \in C \cap B$. Hence, $c \rightarrow_B u$. With u having been

chosen arbitrarily, we actually have that for all $u \in U \cap B$, there is some $c \in C \cap B$ with $c \rightarrow_B u$. It remains to be shown that there is some $c \in C \cap TEQ(B)$ with $c \rightarrow_B^* u$.

To this end, again consider $c_i \in C \cap B$ such that $\overline{D}_B(c_i) \cap C = \emptyset$. It suffices to show that $c_i \rightarrow_B^* b$ for all $b \in B$, as then both $c_i \in TEQ(B) \cap C$ and $c_i \rightarrow_B^* u$. So, consider an arbitrary $b \in B$. If $b = c_i$, the case is trivial. If $b \in C \cap B$ but $b \neq c_i$, we are done by (1). If instead $b \in U \cap B$, then $c \rightarrow_B^* b$ for some $c \in C \cap B$, as we have shown in the first part of the proof. If $c = c_i$, we are done. Otherwise, we can apply (1) to obtain $c_i \rightarrow_B c' \rightarrow_B^* b$ and hence $c_i \rightarrow_B^* b$. □

We are now ready to state the main theorem of this paper.

Theorem 2 *Deciding whether a particular alternative is in the tournament equilibrium set of a tournament is NP-hard.*

Proof: By reduction from *3SAT*. Consider an arbitrary *3CNF* φ and construct the tournament $T_\varphi^{TEQ} = (C \cup U, \succ)$. This can be done in polynomial time. We show that

$$\varphi \text{ is satisfiable if and only if } d \in TEQ(T_\varphi^{TEQ}).$$

For the direction from left to right, observe that by an argument analogous to the proof of Theorem 1 it can be shown that φ is satisfiable if and only if $d \in BA(T_\varphi^{TEQ})$. So assuming that φ is not satisfiable yields $d \notin BA(T_\varphi^{TEQ})$. By the inclusion of TEQ in the Banks set (Proposition 1), it follows that $d \notin TEQ(T_\varphi^{TEQ})$.

For the opposite direction, assume that φ is satisfiable. Then there is a choice set W for $\{X_i\}_{1 \leq i \leq m}$ such that $x' = \bar{x}$ for no $x, x' \in W$. Obviously $W \cup \{y_1, \ldots, y_{m-1}\} \cup \{z_i^j \in Z \colon x_i^j \in W\} = \{u_1, \ldots, u_n\}$ contains no cycles and thus is transitive. Without loss of generality we may assume that $u_i \in U_i$ for all $1 \leq i \leq n$. For each $1 \leq i \leq n+1$, define a subset $\overline{D}_i$ of alternatives as follows. Set $\overline{D}_{n+1} = A$ and $\overline{D}_i = \bigcap_{i \leq j \leq n+1} \overline{D}(u_j)$ for each $1 \leq i \leq n$. Hence, $\overline{D}_1 \subsetneq \cdots \subsetneq \overline{D}_{n+1}$. In an effort to simplify notation, we write $\rightarrow_i$ and $\overline{D}_i(x)$ for $\rightarrow_{\overline{D}_i}$ and $\overline{D}_{\overline{D}_i}(x)$, respectively. It then suffices to prove that

$$d \in TEQ(\overline{D}_k), \text{ for all } 1 \leq k \leq n+1. \tag{2}$$

The theorem then follows as the special case in which $k = n+1$. We first make the following observations concerning the TEQ relation $\rightarrow_i$ in each $\overline{D}_i$, for each $1 \leq i, j \leq n+1$:

(*i*) $u_j \in \overline{D}_i$ if and only if $j < i$,

(*ii*) $c_j \in \overline{D}_i$ if and only if $j < i$,

(*iii*) $c_i \rightarrow_{i+1} c_j$ if $j < i \leq n$,

(*iv*) $u_i \rightarrow_{i+1} c_i$, if $i \leq n$.

For (*i*), observe that if $j < i$, $u_j \in \overline{D}(u_i)$ by transitivity of the set $\{u_1, \ldots, u_n\}$. Hence, $u_j \in \overline{D}_i$. If on the other hand $j \geq i$, then $u_j \notin \overline{D}(u_j)$ and thus $u_j \notin \overline{D}_i$. For (*ii*), observe that $c_j \in \overline{D}(u_i)$ for all $i \neq j$ and thus $c_j \in \overline{D}_i$ if $j < i$. However, $c_j \notin \overline{D}(u_j)$ and hence $c_j \notin \overline{D}_i$ if $j \geq i$. For (*iii*), merely observe that c_i is the Condorcet winner in $\overline{D}_{i+1}(c_j)$, if $j < i \leq n$. To appreciate (*iv*), observe that by construction $\overline{D}_{i+1}(c_i)$ has to be either a singleton $\{u_i\}$ for some $u_i \in U_i$, or U_i itself. The former is the case if $U_i \subseteq Y$, or if $U_i \subseteq X$ and $i \neq n$. The latter holds if $U_i = U_n$ or if $U_i \subseteq Z$. In either case, $TEQ(\overline{D}_{i+1}(c_i)) = \overline{D}_{i+1}(c_i)$ and $u_i \rightarrow_{i+1} c_i$ holds. For the case in which $U_i \subseteq X$ with $i \neq n$, let $U_i = \{u_i, u_i', u_i''\}$. By construction, $U_{i+2} \subseteq Z$ and $u_i', u_i'' \notin \overline{D}(u_{i+2})$. Accordingly, $u_i', u_i'' \notin \overline{D}_{i+1}$. From $u_i \in \overline{D}_{i+1}$ it then follows that $\overline{D}_{i+1} \cap U_i = \{u_i\}$.

Algorithm 1 Tournament Equilibrium Set

procedure $\texttt{TEQ}(X)$
$R \leftarrow \emptyset$
$B \leftarrow C \leftarrow \arg\min_{a \in X} |\overline{D}(a)|$
loop
  $R \leftarrow R \cup \{(b, a) \colon a \in C \wedge b \in \texttt{TEQ}(\overline{D}(a))\}$
  $D \leftarrow \bigcup_{a \in C} \texttt{TEQ}(\overline{D}(a))$
  if $D \subseteq B$ **then return** $TC_B(R)$ **end if**
  $C \leftarrow D$
  $B \leftarrow B \cup C$
end loop

We are now in a position to prove (2) by induction on k. For $k = 1$, observe that d is a Condorcet winner in $\overline{D}_1$ and thus $d \in \mathit{TEQ}(\overline{D}_1)$. For the induction step, let $k = i + 1$. With observation (i) we know that $u_i \in \overline{D}_{i+1}$ and, in virtue of the induction hypothesis, also that $d \in \mathit{TEQ}(\overline{D}_i)$. Hence, $d \rightarrow_{i+1} u_i$. Moreover, by observations (iii) and (iv), $c_i \rightarrow_{i+1} d \rightarrow_{i+1} u_i \rightarrow_{i+1} c_i$, *i.e.*, c_i, d and u_i constitute a $\rightarrow_{i+1}$-cycle. In virtue of Lemma 1 and observation (ii), we may conclude that $c_i \rightarrow^*_{i+1} a$ for all $a \in \overline{D}_{i+1}$. Accordingly, $\{c_i, d, u_i\} \subseteq TC_{\overline{D}_{i+1}}(\rightarrow_{i+1})$ and $d \in \mathit{TEQ}(\overline{D}_{i+1})$, which concludes the proof. □

5 A Heuristic for Computing TEQ

Computational intractability of the TEQ *membership* problem implies that TEQ cannot be computed efficiently either. Nevertheless, the running time of the naive algorithm, which straightforwardly implements the recursive definition of TEQ, can be greatly reduced when assuming that TEQ satisfies CTC. This assumption can fairly be made. For if CTC were not to hold, TEQ would be non-monotonic and thus compromised as a solution concept, the issue of computing it moot.

Algorithm 1 computes TEQ by starting with the set B of all alternatives that have dominator sets of minimal size (*i.e.*, the so-called Copeland winners). These alternatives are good candidates to be included in TEQ and the small size of their dominator sets speeds up the computation of their TEQ-dominators. Then, all alternatives that TEQ-dominate any alternative in B are iteratively added to B until no more such alternatives can be found, in which case the algorithm returns the top cycle of $\rightarrow_B$. Of course, the *worst-case* running time of this algorithm is still exponential, but experimental results suggest that it outperforms the naive algorithm by a factor of about five in uniform random tournaments with up to 150 vertices (see Table 1). We implemented two versions of the naive algorithm, which differ in the subroutine that determines the top cycle. The first one uses the Floyd-Warshall algorithm with an asymptotic complexity of $O(n^3)$, whereas the second one employs Kosaraju's algorithm with a complexity of $O(n^2)$ (see, *e.g.*, Cormen et al., 2001). Surprisingly, the variant relying on Floyd-Warshall performs slightly better on moderately sized instances due to factors hidden in the asymptotic notation that are amplified as a consequence of TEQ's recursive definition.

While choosing tournaments uniformly at random might be useful for benchmarking algorithms, it raises a number of conceptual problems. First, in voting and most other applications uniform random tournaments do not represent a reasonably realistic model of social preferences. Secondly, these tournaments are "almost" regular and tournament solutions almost always select all alternatives in regular tournaments. One model of random tournaments that have more structure can be obtained by defining an arbitrary linear order

$\|A\|$	Floyd-Warshall	Kosaraju	Algorithm 1
Uniform random tournaments ($p = 0.5$)			
50	0.48 s	0.59 s	0.09 s
100	53.33 s	65.73 s	9.57 s
150	1 166 s	1 429 s	210 s
Structured random tournaments ($p = 0.8$)			
50	13.87 s	16.56 s	0.01 s
100	18 416 s	21 382 s	8.46 s
150	—	—	1273 s

Table 1: Experimental evaluation of algorithms that compute TEQ. Average running time for ten instances on a 3.2GHz Core2Duo machine. Both versions of the naive algorithm did not terminate within 24 hours when run on structured random tournaments with 150 vertices.

on the alternatives $a_1, \ldots, a_m$ and letting $a_i \succ a_j$ for $i < j$ with probability $p > 0.5$. Letting $p = 1$ yields a "completely structured" transitive tournament. The more structure a tournament possesses, the more Algorithm 1 outperforms the naive algorithm, due to the increasing number of large dominator sets that have to be analyzed by the latter at every level of the recursion. In large structured tournaments, the performance gap becomes rather impressive (see Table 1). For example, the naive algorithm requires more than five hours to compute the TEQ of a structured random tournament with 100 vertices whereas it takes Algorithm 1 about eight seconds.[1]

We have further used the naive algorithm to try to disprove CTC (and thus TEQ's monotonicity), but failed to find a counterexample by an exhaustive search in all tournaments with up to ten vertices (roughly ten million non-isomorphic tournaments), all regular tournaments with up to 13 vertices, and all locally transitive tournaments with up to 20 vertices. We also investigated a fairly large number of uniform and structured random tournaments, again to no avail. This can be considered mild evidence that TEQ is indeed monotonic.

Acknowledgements

We are grateful to Brendan McKay for providing extensive lists of non-isomorphic tournaments. This material is based upon work supported by the Deutsche Forschungsgemeinschaft under grant BR 2312/3-2.

References

N. Alon. Ranking tournaments. *SIAM Journal of Discrete Mathematics*, 20(1):137–142, 2006.

D. Bouyssou, T. Marchant, M. Pirlot, A. Tsoukiàs, and P. Vincke. *Evaluation and Decision Models: Stepping Stones for the Analyst*. Springer-Verlag, 2006.

F. Brandt and F. Fischer. Computing the minimal covering set. *Mathematical Social Sciences*, 2008. To Appear.

[1] We also had some limited success with algorithms that make use of the easy-to-prove fact that $TEQ((A, \succ)) = TEQ(TC(\succ))$. Assuming CTC, a similar preprocessing step that first computes the minimal covering set of the tournament at hand is possible.

F. Brandt and P. Harrenstein. Dominance in social choice and coalitional game theory. In G. Bonanno, B. Löwe, and W. van der Hoek, editors, *Proceedings of the 8th Conference on Logic and the Foundations of Game and Decision Theory (LOFT)*, 2008.

F. Brandt, F. Fischer, and P. Harrenstein. The computational complexity of choice sets. In D. Samet, editor, *Proceedings of the 11th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pages 82–91. ACM Press, 2007.

V. Conitzer. Computing Slater rankings using similarities among candidates. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, pages 613–619. AAAI Press, 2006.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 2nd edition, 2001.

Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité de décisions rendues à la pluralité de voix*. Imprimerie Royal, 1785. Facsimile published in 1972 by Chelsea Publishing Company, New York.

J. Duggan and M. Le Breton. Dutta's minimal covering set and Shapley's saddles. *Journal of Economic Theory*, 70:257–265, 1996.

P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

P. E. Dunne. Computational properties of argumentation systems satisfying graph-theoretic constraints. *Artificial Intelligence*, 171(10-15):701–729, 2007.

B. Dutta. On the tournament equilibrium set. *Social Choice and Welfare*, 7(4):381–383, 1990.

B. Dutta and J.-F. Laslier. Comparison functions and choice correspondences. *Social Choice and Welfare*, 16(4):513–532, 1999.

P. C. Fishburn. Condorcet social choice functions. *SIAM Journal on Applied Mathematics*, 33(3):469–489, 1977.

D. B. Gillies. Solutions to general non-zero-sum games. In A. W. Tucker and R. D. Luce, editors, *Contributions to the Theory of Games IV*, volume 40 of *Annals of Mathematics Studies*, pages 47–85. Princeton University Press, 1959.

G. Laffond, J.-F. Laslier, and M. Le Breton. More on the tournament equilibrium set. *Mathématiques et sciences humaines*, 31(123):37–44, 1993.

J.-F. Laslier. *Tournament Solutions and Majority Voting*. Springer, 1997.

C. H. Papadimitriou. *Computational Complexity*. Addison-Wesley, 1994.

T. Schwartz. Cyclic tournaments and cooperative majority voting: A solution. *Social Choice and Welfare*, 7:19–29, 1990.

G. J. Woeginger. Banks winners in tournaments are difficult to recognize. *Social Choice and Welfare*, 20:523–528, 2003.

Felix Brandt, Felix Fischer, and Paul Harrenstein
Institut für Informatik
Ludwig-Maximilians-Universität München
80538 Munich, Germany
Email: {brandtf,fischerf,harrenst}@tcs.ifi.lmu.de

Maximilian Mair
Mathematisches Institut
Ludwig-Maximilians-Universität München
80538 Munich, Germany
Email: mairm@cip.ifi.lmu.de

Approximability of Manipulating Elections[1]

Eric Brelsford, Piotr Faliszewski, Edith Hemaspaandra,
Henning Schnoor and Ilka Schnoor

Abstract

In this paper, we set up a framework to study approximation of manipulation, control, and bribery in elections. We show existence of approximation algorithms (even fully polynomial-time approximation schemes) as well as obtain inapproximability results. In particular, we show that a large subclass of scoring protocols admits fully polynomial-time approximation schemes for the coalitional weighted manipulation problem and that if certain families of scoring protocols (e.g., veto) admitted such approximation schemes then P = NP. We also show that bribery for Borda count is NP-complete and that there is no approximation algorithm that achieves even a polynomial approximation ratio for bribery in Borda count for the case where voters have prices.

1 Introduction

Elections are an essential mechanism that each democratic society uses to make joint decisions. They are also important tools within computer science. For example, [DKNS01] show how to build a meta search-engine via conducting elections between other search engines; Ephrati and Rosenschein [ER97] use voting to solve certain planning problems; and in the context of multiagent systems, elections and voting are naturally used to obtain the joint decisions of agent societies.

Unfortunately, a famous result of Gibbard [Gib73] and Satterthwaite [Sat75] states that (see also the work of Duggan and Schwartz [DS00]) for any reasonable election system (with at least 3 candidates) there exist scenarios where at least some voters have an incentive to vote strategically, i.e., to vote not according to their true preferences but in a way that yields a result more desirable for them. Similarly, the result of an election can be skewed by an external agent who bribes some of the voters to change their votes or even by the authority organizing the election, via, e.g., encouraging or discouraging particular candidates from participating, or via arranging voting districts in a certain way.

The possibility that the result of the election can be skewed via strategic voting, bribery, or procedural control is very disconcerting. In the early 90s, Bartholdi, Orlin, Tovey, and Trick [BTT89, BO91, BTT92] suggested a brilliant way to circumvent this issue. They observed that since all voters are computationally-bounded entities, even if various forms of manipulating elections are possible in principle,[2] they constitute a real threat only if it is computationally easy, for a given election system, to determine the appropriate actions that affect the result in the desired way (i.e., for the case of strategic voting to determine how the manipulators should vote; for the case of bribery determine who to bribe and how, etc.). To measure the computational difficulty of manipulation and control, Bartholdi, Orlin, Tovey, and Trick used the complexity-theoretic notion of NP-hardness.

The ideas of Bartholdi, Orlin, Tovey and Trick did not receive that much attention until a few years ago, when it became apparent that elections and voting are important tools in the context of multiagent systems, and that software agents are capable of much more systematic attempts at manipulating elections than, say, humans. Thus, in recent

[1] This is an extended version of the AAAI-08 paper with the same title.

[2] In the literature the technical term *manipulation* means strategic voting. In this section by *manipulating* we mean the general notion of affecting the result of an election.

years many papers focused on the computational analysis of voting rules with respect to manipulation, e.g., [CSL07, EL05, PR07, PRZ07, HH07], bribery [FHH06, FHHR07, Fal08], and control [HHR07, FHHR07, PRZ07]. Most of these papers focus on obtaining polynomial-time algorithms and NP-hardness results for various forms of affecting the result of elections. However, NP-hardness gives only worst-case complexity guarantees, and it might very well be the case that even though, say, manipulation in a given election system is NP-hard, finding effective manipulations is often easy. Recently, Conitzer and Sandholm [CS06], and Procaccia and Rosenschein [PR07] looked into these issues and they provide examples of voting rules and distributions of votes for which this is the case. We will refer to the approach presented, among others, in these two papers as *frequency of hardness approach*.

In this paper we refine the study of the complexity of manipulating elections via analysis of approximability of NP-hard election-manipulation problems. An important contribution of this paper is a natural, uniform objective function that can be used to measure the effectiveness of a particular manipulation, bribery, or control attempt. Thus we set up a general framework for studying approximation for these problems.[3] Our function is particularly interesting for the case of manipulation, where defining a natural objective function is not straightforward.

We show existence of approximation algorithms (even fully polynomial-time approximation schemes) for manipulation for a large subclass of voting rules known as scoring protocols (for bounded numbers of candidates) as well as obtain inapproximability results regarding several prominent families of scoring protocols (e.g., veto and k-approval for unbounded number of candidates). We prove NP-hardness of bribery for Borda count and inapproximability of bribery for Borda count for the case where each voter has a price for changing its vote. Hardness results for control (i.e., changing the outcome of the election by *adding* voters) of unweighted Borda elections have been obtained by Russell [Rus07]. We use a similar technique to prove that the bribery problem for unweighted Borda is NP-complete. To the best of our knowledge, our NP-hardness result for Borda is the first hardness result for a problem of affecting the result of unweighted Borda count elections via *modifying* the voters.

Related work Several previous papers studied approximation of various manipulation and bribery problems but each of them used objective functions specifically tailored to their tasks. In particular, Faliszewski [Fal08] studied approximability of the total cost of a bribery for plurality and approval voting. Zuckerman, Procaccia and Rosenschein [ZPR08], among other things, studied approximability of the minimum number of unweighted manipulators needed to change the result of an election for several voting systems, including Borda count.

We contrast our approach and results with the frequency of hardness approach. The existence of an approximation algorithm (in particular, the existence of a fully polynomial-time approximation scheme) for a given election problem is much stronger evidence that this problem is practically easy than a frequency of hardness result stating that the problem is easy often, according to some distribution. The reason for this is that a polynomial-time approximation algorithm guarantees to find a near-optimal answer for *every* input instance. If our problem is frequently easy it might still be the case that the instances that we encounter in practice happen to be the "rare" difficult ones. On the other hand, inapproximability is a worst case notion. If a problem is in general inapproximable, it might still be the case that most of its instances are easy (are easily approximable). Nonetheless, inapproximability of a given NP-hard election problem is stronger evidence of its computational hardness than NP-hardness alone.

We focus on manipulation and bribery rather than on control, but we mention that Brelsford [Bre07] studied several control problems from the point of view of approximation.

[3]To be technically correct, our approach is limited to voting rules that assign numerical scores to the candidates. This is the case for most standard voting rules.

2 Preliminaries

Elections An election E is a pair (C, V), where C is a finite set of candidates and V a finite multiset of strict linear orders on C. An order $v \in V$ is called a vote and represents the preference of a voter over the candidates. The winner of an election E depends on the underlying election system. In this paper we consider only election systems that are represented by scoring protocols and families of scoring protocols. A scoring protocol is a vector $(\alpha_1, \ldots, \alpha_m)$ of natural numbers with $\alpha_1 \geq \cdots \geq \alpha_m$. Using this protocol the winner of an election E with m candidates can be determined as follows: Every candidate c gets α_i points for every vote that ranks c in the ith place and $\text{score}_E(c)$ is the sum of all points c gets in this way. In the end c is a winner if no other candidate has a higher score. We also allow the votes in E to have weights, in this case each vote with weight $w \in \mathbb{N}$ is counted as w identical votes.

Let $(S_i)_{i\geq 1}$ be a family of scoring protocols such that S_i is a scoring protocol of length i. We represent by $(S_i)_{i\geq 1}$ the election system that uses S_m to determine the winner of an election with m candidates. Borda count is the election system using $((i-1, i-2, \ldots, 0))_{i\geq 1}$, and veto is the election system using $((1, 1, \ldots, 1, 0))_{i\geq 1}$.

Approximating Elections In this paper we study approximation algorithms for manipulation and bribery. In both problems our goal is to ensure that a specified candidate is a winner but in manipulation we attempt to reach this goal via, in essence, adding a certain number of votes, whereas in bribery we do so via changing up to a given number of votes. (Note that sometimes manipulation is defined as allowing to *change* a *specified* set of voters. Our version allows to state our results in an easier notation—it is easy to see that these notions describe the exact same issue. One may view adding the manipulators' votes as a process where the manipulators make up their minds as to how to vote, and then cast their votes. Note that unlike for e.g., bribery, the original votes of the manipulators are completely irrelevant for the problem to determine if a designated candidate can be made a winner.) We also study the manipulation problem where the voters additionally have weights, and the bribery problem where the votes have prices which the briber has to pay in order to change the vote.

We require our approximation algorithms to produce "solutions" to their respective instances. A solution is a strategy specifying which actions to perform, i.e., what votes to add for the case of manipulation and which votes to change (and how to change them) for the case of bribery.

In our model we assume that we know all the votes that are supposed to be cast in the election. In reality, however, we are often limited to only having a guess regarding these votes. Thus we are interested in finding a strategy that benefits the specified candidate as much as possible so that this candidate has a good change of becoming a winner even if the guess is a little off.

In the setting of scoring protocols (or any other score-based election system), a natural way to measure the performance of a candidate p in an election E (written as $\text{perf}^E(p)$) is $\text{score}_E(p) - \max\{\text{score}_E(c) \mid c \in C \setminus \{p\}\}$, the difference between the score of p and that of p's strongest competitor. $\text{perf}^E(p)$ tells us "how much" p wins the election or "how close" p is to winning it. Obviously, p wins the election E if and only if $\text{perf}^E(p)$ is nonnegative.

A natural measure of the effectiveness of a manipulating action s within election E is the increase of performance of the favorite candidate p obtained by applying this action.

Definition 1 *$\beta(E, s) = \text{perf}^{s(E)}(p) - \text{perf}^E(p)$, where $s(E)$ denotes the election resulting from applying action s to E.*

Note that the β function allows us to deal with uncertainty in a natural way: If we only have knowledge about a part of the election, then it is a natural goal to give our candidate as

much of a headstart in the part of the election that we do know as possible. This is exactly what is expressed in the β-function. Also, β can be applied not only to manpulation and bribery as studied in this paper, but to just about every possible way to interfere with the result of an election. We therefore believe it to be a uniform way to describe the "success" of the action of a dishonest party in an election scheme.

Finally observe that for given strategies (added voters, bribes, etc.), the value of β can be negative. However, for scoring protocols (and most other natural election rules) it is easy to come up with strategies that have a nonnegative value of β (strategies that do nothing at all suffice). Hence we require our approximation algorithms to only output "reasonable" strategies, i.e., strategies for which the value of β is nonnegative. We now define our optimization problems (which we will prefix with the election system under consideration):

\$-bribery-max The input I consists of an election E, for each existing vote a natural number defining its price, a preferred candidate p, and a natural number k representing the budget available to the briber. Solutions consist of a set of votes in the election E such that the sum of their prices does not exceed the budget k, and new votes to replace them with. The goal is to find a solution s maximizing $\beta(E, s)$.

weighted-manipulation-max Here, the input I consists of an election E where each vote is accompanied by its weight, the preferred candidate p, and a list of weights (of the manipulators). A solution consists of a vote for each manipulator. Again, the goal is to find a solution s such that $\beta(E, s)$ is maximal.

Note that if we could compute the maximum value of β for a given optimization problem then, naturally, we could solve the corresponding decision problem. This means that if the corresponding decision problem is NP-hard (as is often the case for manipulation and bribery) then we cannot hope to compute the optimal value of β exactly. However, since β is defined as an increase in p's performance and thus its maximum value is positive in most settings, we can attempt to use natural techniques to compute it approximately.

Approximation Algorithms and Elections The quality of an approximation algorithm is usually measured by comparing the solutions it computes to the optimal ones. For our optimization problems, an instance I contains the election itself, the preferred candidate, and additional parameters limiting the possible strategies for affecting the result of the election (i.e., the available budget in \$-bribery and the weights of the manipulators in manipulation). For such an instance I containing the election E, we define an *optimal solution* to be a solution s that achieves the maximal possible value of $\beta(E, s)$ among *all* legal solutions. We define $\mathrm{OPT}(I)$ as $\beta(E, s)$ for such an optimal solution s.

Given an instance I, an approximation algorithm $\mathcal{A}$ is required to produce a legal solution $\mathcal{A}(I)$, that is, a solution that respects the constraints specified in I. For a positive real constant c, we say that $\mathcal{A}$ is a *factor c approximation algorithm*, if for each instance I containing the election E, we have that $\beta(E, \mathcal{A}(I)) \geq \frac{1}{c}\mathrm{OPT}(I)$. Such an algorithm guarantees that the effectiveness of the solution obtained from applying $\mathcal{A}$ to the instance is at least $\frac{1}{c}$ of the effectiveness that the optimal solution achieves.

We say that an algorithm $\mathcal{A}$ is a *polynomial-time approximation scheme* if for each input (I, ε), where ε is a rational value between 0 and 1 and where I contains an election E, it holds that: (a) $\mathcal{A}$ produces a solution $s = \mathcal{A}(I, \varepsilon)$ such that $\beta(E, s) \geq (1-\varepsilon)\cdot\mathrm{OPT}(I)$, and (b) for each fixed value of ε, $\mathcal{A}$ runs in time polynomial in $|I|$. If, in fact, $\mathcal{A}$ runs in time polynomial in both $|I|$ and $\frac{1}{\varepsilon}$ then $\mathcal{A}$ is a *fully polynomial-time approximation scheme* (FPTAS). In the current paper we only consider maximization problems, analogous definitions can be given for minimization problems as well.

3 Manipulation in Scoring Protocols

Hemaspaandra and Hemaspaandra [HH07] showed that for each scoring protocol $\alpha = (\alpha_1, \ldots, \alpha_m)$ such that it is not the case that $\alpha_2 = \cdots = \alpha_m$ the problem α-weighted-manipulation is NP-complete (see also [CSL07, PR07]). In this section we show that, nonetheless, weighted manipulation is easy for a large class of scoring protocols in practice. We do so via showing FPTASes for the scoring protocols in this class.

Let α be a scoring protocol. An instance I of α-weighted-manipulation-max is a tuple (E, w, p) where $E = (C, V)$ is an election with candidate set C and weighted nonmanipulative voters V, $w = (w_1, \ldots, w_n)$ is a sequence of weights of the manipulative voters, and $p \in C$ is our preferred candidate. Our goal is to maximize the performance of p. That is, our goal is to find a solution *sol* such that $\beta(E, sol) = \mathrm{OPT}(I)$.

Theorem 2 *Let $\alpha = (\alpha_0, \ldots, \alpha_m)$ be a scoring protocol such that $\alpha_0 > \alpha_1$. There is an algorithm $\mathcal{A}$ that given a rational number ε, $0 < \varepsilon < 1$, and an instance $I = (E, w, p)$ of α-weighted-manipulation-max computes, in polynomial time in $|I|$ and $\frac{1}{\varepsilon}$, a solution sol such that $\beta(E, sol) \geq (1 - \varepsilon)\mathrm{OPT}(I)$.*

Note that Theorem 2 claims that for each *separate* scoring protocol $(\alpha_0, \ldots, \alpha_m)$, where $\alpha_0 > \alpha_1$, there is a *separate* algorithm. In particular, each of the algorithms from Theorem 2 is tailored for a fixed number of candidates. Before we jump to the proof, we need to introduce some notation.

Let $\alpha = (\alpha_0, \ldots, \alpha_m)$ be a scoring protocol where $\alpha_0 > \alpha_1$ and let $C = \{p, c_1, \ldots, c_m\}$ be a set of candidates. p is our preferred candidate whose performance we want to maximize. We implicitly assume that we have a set V of nonmanipulative voters, however in this discussion the only incarnation of the nonmanipulative voters is through the sequence s below.

We let $w = (w_1, \ldots, w_n)$ be the sequence of weights of the manipulators. Naturally, to maximize p's performance, each manipulator ranks p first. The complexity of α-weighted-manipulation-max comes from the difficulty in arranging the remainder of the manipulators' votes in such a way as to minimize the score of p's most dangerous competitor.

By $\mathcal{E}(C, w)$ we mean the set of all elections over the candidate set C with voter set containing exactly voters with weights $w_1, \ldots, w_n$. Let $s = (s_1, \ldots, s_m)$ be a sequence of nonnegative integers. Intuitively, the sequence s gives the scores that candidates c_1 through c_m receive from the nonmanipulative voters. By $S_\alpha(E, s)$ we mean $\max_{i \in \{1,\ldots,m\}}\{\mathrm{score}_E(c_i) + s_i\}$ and by $T_\alpha(w, s)$ we mean $\min_{E \in \mathcal{E}(C,w)} S_\alpha(E, s)$. Function $T_\alpha(w, s)$ measures the smallest possible top score of a candidate from $\{c_1, \ldots, c_m\}$ after the manipulators cast their votes. We now prove that for each scoring protocol α there is an FPTAS for T_α.

Lemma 3 *Let $\alpha = (\alpha_0, \ldots, \alpha_m)$ be a scoring protocol and let $C = \{p, c_1, \ldots, c_m\}$. There is an algorithm $\mathcal{T}$ that given a rational number ε, $0 < \varepsilon < 1$, a sequence $s = (s_1, \ldots, s_m)$ of nonnegative integers and a sequence of manipulators weights $w = (w_1, \ldots, w_n)$ computes an election $E \in \mathcal{E}(C, w)$ such that $S_\alpha(E, s) \leq (1 + \varepsilon)T_\alpha(w, s)$. Algorithm $\mathcal{T}$ runs in polynomial time in n, m, and $\frac{1}{\varepsilon}$.*

Proof. Set $w_{\max} = \max\{w_1, \ldots, w_n\}$ and set $K = \frac{\varepsilon w_{\max}}{n\alpha_1}$. Set $w' = (K\lceil\frac{w_1}{K}\rceil, \ldots, K\lceil\frac{w_n}{K}\rceil)$. It is possible to compute in polynomial time in n, m, and $\frac{1}{\varepsilon}$ an election $E' \in \mathcal{E}(C, w')$ such that $S_\alpha(E', s) = T_\alpha(w', s)$. (One can do so via a routine dynamic programming approach; we enforce that in our solution each voter ranks p first.) Let E be an election identical to E' only that appropriate voters have weights $w_1, \ldots, w_n$ instead of $w'_1, \ldots, w'_n$. Our algorithm outputs E.

It is easy to see that our algorithm can be made to work in polynomial time as required. Let us now show that the solution it produces satisfies the requirements regarding quality.

It is easy to see that $T_\alpha(w,s) \geq \alpha_1 w_{\max}$ and that $S_\alpha(E',s) \leq T_\alpha(w,s) + \alpha_1 nK$. The former is true because some candidate needs to get α_1 points from the manipulator with weight $w_{\max}$ and the second follows from the fact that for each i in $\{1,\ldots,n\}$ we have $w_i \leq w'_i < w_i + K$. For the same reason $S_\alpha(E,s) \leq S_\alpha(E',s)$.

Thus, $S_\alpha(E,s) \leq T_\alpha(w,s) + \alpha_1 nK = T_\alpha(w,s) + \varepsilon w_{\max}$. Since $T_\alpha(w,s) \geq \alpha_1 w_{\max}$, this yields that $S_\alpha(E,s) \leq (1+\varepsilon)T_\alpha(w,s)$. (Note that, technically, this argument is only correct if $\alpha_1 \geq 1$ but, naturally, if $\alpha_1 = 0$ then the theorem is trivially satisfied.) This completes the proof. □

With Lemma 3 at hand we can prove Theorem 2.

Proof. Our input is $I = (E,w,p)$, where $E = (C,V)$ is an election with candidate set $C = \{p, c_1 \ldots, c_m\}$ and set V of nonmanipulative voters, $w = (w_1,\ldots,w_n)$ is a sequence of manipulators' weights, and p is our preferred candidate. Our goal is to find a solution *sol* (a collection of votes for the manipulators to cast) that maximizes $\beta(E, sol)$.

Let $W = \sum_{i=1}^{n} w_i$ and let $w_{\max} = \max\{w_1,\ldots,w_n\}$. For each i in $\{1,\ldots,m\}$ let $s_i = \mathrm{score}_E(c_i)$. We assume that the candidates $c_1,\ldots,c_m$ are listed in such an order that $s_1 \geq s_2 \geq \cdots \geq s_m$. Since $\alpha_0 > \alpha_1$, in every optimal solution each manipulator ranks p first and so $\mathrm{OPT}(I) = W\alpha_0 - (T_\alpha(w,s) - s_1)$. It would seem that computing approximately $T_\alpha(w,s)$ should be enough to get a good approximation of $\mathrm{OPT}(I)$, but unfortunately $T_\alpha(w,s)$ can be much bigger than $\mathrm{OPT}(I)$. We have to, in some sense, reduce its value first.

Note that we can disregard all candidates c_j such that $s_1 - s_j > \alpha_1 W$. If there are k such candidates then the manipulators may simply rank them on the first k positions after p. For the sake of simplicity, we assume that there are no such candidates.

Let $s' = (s_1 - s_m, \ldots, s_m - s_m)$. It is easy to see that $\mathrm{OPT}(I) = W\alpha_0 - (T_\alpha(w,s) - s_1) = W\alpha_0 - (T_\alpha(w,s') - s'_1)$. Additionally, via the above paragraph, we have that for each s'_i it holds that $s'_i \leq \alpha_1 W$. However, this means that $T_\alpha(w,s') \leq 2\alpha_1 W$. This is so because at worst the candidate whose score is the value of $T_\alpha(w,s')$ gets $\alpha_1 W$ points from s' and another $\alpha_1 W$ points from the manipulators.

Using algorithm $\mathcal{T}$ from Lemma 3, we fill-in the manipulators' votes to form an election $E' \in \mathcal{E}(C,w)$ such that all voters in E' rank p first and $T_\alpha(w,s') \leq S_\alpha(s',E') \leq (1+\varepsilon')T_\alpha(w,s')$, where $\varepsilon' = \frac{1}{2\alpha_1}\varepsilon$. (Recall that in our setting α_1 is a constant.) Votes obtained in this way are the solution *sol* that our algorithm produces and we have $\beta(E,sol) = W\alpha_0 - (S_\alpha(s',E') - s'_1)$. Note that

$$\begin{aligned}
\mathrm{OPT}(I) &= W\alpha_0 - (T_\alpha(w,s') - s'_1) \\
&\geq W\alpha_0 - (S_\alpha(s',E') - s'_1) \\
&\geq W\alpha_0 - ((1+\varepsilon')T_\alpha(w,s') - s'_1) \\
&= W\alpha_0 - (T_\alpha(w,s') - s'_1) - \varepsilon' T_\alpha(w,s') \\
&= \mathrm{OPT}(I) - \varepsilon' T_\alpha(w,s').
\end{aligned}$$

Since $\mathrm{OPT}(I) \geq W$ (this is a consequence of the fact that $\alpha_0 > \alpha_1$), $T_\alpha(w,s') \leq 2\alpha_1 W$, and $\varepsilon' = \frac{1}{2\alpha_1}\varepsilon$, via the above calculations, $\mathrm{OPT}(I) \geq W\alpha_0 - (S_\alpha(s',E') - s'_1) \geq (1-\varepsilon)\mathrm{OPT}(I)$ and thus $\mathrm{OPT}(I) \geq \beta(E,sol) \geq (1-\varepsilon)\mathrm{OPT}(I)$. This completes the proof. □

Interestingly, we can use Theorem 2 to obtain results similar to those of Zuckerman, Procaccia, and Rosenschein [ZPR08], but for the case of scoring protocols $\alpha = (\alpha_0,\ldots,\alpha_m)$ such that $\alpha_0 > \alpha_1$. Note that Theorem 4 below says that there is a *separate* algorithm for *each* scoring protocol of the given form. (Also, keep in mind that each *single* scoring protocol only works with a fixed number of candidates.)

Theorem 4 *Let ε be a rational number, $0 < \varepsilon < 1$, and let $\alpha = (\alpha_0, \ldots, \alpha_m)$ be a scoring protocol such that $\alpha_0 > \alpha_1$. There is an algorithm that given an instance $I = (E, w, p)$ of α-weighted-manipulation, where $w = (w_1, \ldots, w_n)$ is the sequence of manipulators' weights, has the property that if there is a successful manipulation for instance I, it finds, in polynomial time in $|I|$ and $\frac{1}{\varepsilon}$, a successful manipulation for instance $I' = (E, (w_1, \ldots, w_n, w_{n+1}), p)$, where $w_{n+1} = \lceil \varepsilon \max\{w_1, \ldots, w_n\} \rceil$.*

We omit the easy proof. (The idea is to simply find a good enough approximation and then add a single voter, with appropriate weight, that ranks p first.)

Theorem 2 notwithstanding, we now show that for the case of an unbounded number of candidates there are no FPTASes for veto-weighted-manipulation-max and for k-approval-weighted-manipulation-max, unless P $\neq$ NP.

Theorem 5 *If* P $\neq$ NP*, there is no FPTAS for veto-weighted-manipulation-max.*

To prove Theorem 5 it suffices to show that the unary version of weighted manipulation in veto, i.e., one where each weight is encoded in unary, is NP-complete. In unary-encoded variant of weighted manipulation (in veto and in each fixed scoring protocol) it holds that the maximum value of β function is polynomially bounded. Thus, if there was an FPTAS for veto-weighted-manipulation-max, then one could, via a good enough approximation, solve veto-weighted-manipulation exactly in polynomial time. This is a contradiction if P $\neq$ NP.

Theorem 6 *unary-veto-weighted-manipulation is NP-complete.*

Proof. We will reduce from the NP-complete problem Unary-3-Partition [GJ79]: Given a multiset A of $3m$ positive integers in unary and an integer bound B in unary such that for each $a \in A$, $B/4 < a < B/2$ and such that $\sum_{a \in A} a = mB$, does there exist a partition of A into m subsets $A_1, \ldots, A_m$ such that $\sum_{a \in A_i} a = B$ for all i? (Note that $\|A_i\| = 3$ for all i; hence the problem's name.)

Our reduction works as follows. The election consists of one voter of weight B with preference $c_1 > c_2 > \cdots > c_m > p$ and the manipulators have weights $a_1, \ldots, a_{3m}$.

We claim that there is a successful partition of A if and only if p can be made a winner in our constructed election. First suppose that there exists a partition of A into m subsets $A_1, \ldots, A_m$ such that $\sum_{a \in A_i} a = B$. Let the manipulators corresponding to A_i veto candidate c_i. Note that every candidate c_i receives exactly B vetoes. In the resulting election, $\text{score}(p) = mB$ (p is never vetoed), and $\text{score}(c_i) = B + mB - B = mB$, and so p is a winner of the election. For the converse, suppose the manipulators vote in such a way that p is a winner of the election. Without loss of generality, we may assume that p is never vetoed, and so $\text{score}(p) = mB$. In order for p to be a winner, $\text{score}(c_i)$ can be at most mB. $\sum_{c_i} \text{score}(c_i) = m^2 B$, and so this can only happen if $\text{score}(c_i) = mB$ for all c_i. It follows that each c_i receives exactly B vetoes. Let A_i consist of the multi-set of the weights of the voters that veto c_i. Then $A_1, \ldots, A_m$ is a partition such that $\sum_{a \in A_i} a = B$ for all i. □

The same approach can be used to show NP-hardness (and thus non-existence of FPTAS unless P = NP) for unary manipulation for many other families of scoring protocols.

Theorem 7 *If* P $\neq$ NP*, there is no FPTAS for k-veto-weighted-manipulation-max, k-weighted-approval-manipulation-max, and generalized versions of k-weighted-approval-manipulation-max where, as in k-approval, voters give only points to the first k candidates, but any $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_k > 0$ is allowed.*

4 The Bribery Problem for Borda Count

In this section, we prove hardness results for the Borda count election system.

NP-hardness of the decision version We start by showing that Borda-bribery is NP-complete. In Borda-bribery we are given an election E, a distinguished candidate p, and a nonnegative integer k, and we ask if it is possible to ensure that p is a winner of E via modifying at most k votes. Note that our result regards the simplest variant of bribery where each voter has unit weight and unit price. Hardness of more involved variants (i.e., ones including prices or weights or both) follows naturally.

Our proof works via a reduction from a specifically crafted restriction of the set cover problem.

Problem:	34-XC
Input:	Set S, $\|S\| = n$, sets $T_1, \ldots, T_{\frac{3}{4}n}$, where $\|T_i\| = 4$ for each T_i and where each $s \in S$ is in exactly 3 sets T_i.
Question:	Is there a set $I \subseteq \{1, \ldots, \frac{3}{4}n\}$ such that for $i, j \in I, i \neq j$, $T_i \cap T_j = \emptyset$ and $\bigcup_{i \in I} T_i = S$?

It easily follows from the definition that each correct solution I has exactly $\frac{1}{4}n$ elements. This problem was shown to be NP-complete in [FHS08] (there phrased as a version of 1-in-3-satisfiability which is easily seen to be the same problem).

Theorem 8 *Borda-bribery is* NP*-complete.*

Proof. For a set A of candidates, writing A in a vote means A in some arbitrary, but fixed, order. $\overleftarrow{A}$ denotes the candidates of A in reverse order.

Let $S = \{s_1, \ldots, s_n\}$, $T_1, \ldots, T_{\frac{3}{4}n}$ be an instance of 34-XC. Let $k_1 = n^4$ and $k_2 = n^4 - 8n^3 - 4n^2$ (without loss of generality, assume that n is large enough for k_2 to be positive). Our candidate set C is $\{p\} \cup S \cup P_1 \cup P_2$, where P_1 and P_2 are sets of padding candidates such that $\|P_1\| = k_1$ and $\|P_2\| = k_2$. We set $P = P_1 \cup P_2$. The voter set is defined as follows. For each set T_i, we introduce a voter who votes as follows:

$$v_i = T_i > P_1 > S \setminus T_i > P_2 > p.$$

We also introduce m votes of the form $p > S > P$ and m votes of the form $\overleftarrow{S} > p > \overleftarrow{P}$. By increasing m we increase the point differences between pairs of candidates where one candidate comes from $S \cup \{p\}$ and the other from P, without at the same time affecting the point differences between pairs of candidates where both candidates come from $S \cup \{p\}$ or both come from P. In particular, we can choose m large enough such that with bribing at most $\frac{1}{4}n$ voters, the padding candidates cannot be made to win the election. We choose such a value for m, and hence the briber only has to ensure that the candidate p has at least as many points as each candidate in S in order for p to win the election. This allows us to establish a direct correspondence between bribery attempts bribing at most $\frac{1}{4}n$ voters and the 34-XC instance, by showing that a bribery is successful if and only if the bribed votes of the form v_i correspond to a set cover (and the new votes are set in a reasonable way, i.e., they rank p first, then the padding candidates, and then the candidates in S).

In the $\frac{3}{4}n$ votes v_i introduced earlier, p does not gain any points. For each candidate s_i in S, there are exactly 3 "good votes" (votes corresponding to a set T_j where $s_i \in T_j$), and $\frac{3}{4}n - 3$ "bad votes" (corresponding to sets T_j not containing s_i). In a good vote, s_i gains at least good-min $:= (\|C\| - 4)$ points (since the worst position that s_i can be voted in here is the fourth position). On the other hand, the most points that s_i can make in a good vote is good-max $:= (\|C\| - 1)$ points, this occurs if s_i is in the first position of the vote. For the bad votes, the minimum number of points that s_i can make is bad-min $:= (\|C\| - k_1 - n)$

points (the worst position that s_i can be in for these votes is the $(k_1 + n)$th spot), and the maximum gained in a bad vote is bad-max $:= (\|C\| - k_1 - 5)$ points (the best position to be voted here is the $(k_1 + 5)$th position, since each set T_i contains exactly 4 elements).

Since the briber only has to ensure that p beats the candidates in S, we only need to consider briberies of the form where the "deleted voters" are those corresponding to some set T_i, (deleting votes of this form is obviously better for increasing the performance of p than deleting one of the votes where S and p share the $n+1$ top spots) and the added voters vote p first, then the padding candidates, and then the candidates in S (if the briber wants to make p win, then obviously the bribed voters will vote p first, and since we constructed the election in such a way that the padding candidates cannot win, we can without loss of generality assume that the candidates in S are voted last). We fix such a bribery attempt, and for each element $s_i \in S$, let t_i be the number of good votes for s_i that are deleted by the briber. We show that the bribery is successful if and only if $t_i \geq 1$ for all of the s_i, due to the cardinality restrictions, this then implies that the deleted voters correspond to an exact cover in the sense of 34-XC (since we allow exactly $\frac{1}{4}n$ voters to be bribed).

In order to prove this, we need to show the following: If for an element s_i, the number t_i is at least 1, then the maximal number of points that s_i can have in the bribed election is less than the number of points for the preferred candidate p. On the other hand, if $t_i = 0$, then the minimum number of points that s_i has in the bribed election exceeds the score of the candidate p. We now prove this claim by computing these numbers.

The maximal number of points that s_i can have if $t_i \geq 1$ (obviously, it suffices to consider the case $t_i = 1$) occurs when s_i has the maximum number of possible points in its 2 remaining good votes, and the maximal number of points in its $\frac{1}{2}n - 2$ remaining bad votes. Additionally, the candidate s_i has the above-mentioned M points gained from the votes where all the padding candidates are voted behind all of the candidates in S and the candidate p, and it gains at most $\frac{1}{4}n(n-1)$ points from the additional bribed votes (if the candidate is voted in the first spot of the S-block in each of these votes). Therefore, the maximal number of points in the bribed election for $t_i = 1$ is bribed-max $:= 2 \cdot$ good-max $+$ $(\frac{1}{2}n - 2) \cdot$ bad-max $+ M + \frac{1}{4}n(n-1)$, which is the same as $\frac{3}{4}n^2 + \frac{1}{2}nk_2 - \frac{9}{4}n + 2k_1 + 8 + M$.

On the other hand, for $t_i = 0$, the minimal number of points for a candidate s_i is the following (3 good votes remaining, plus $\frac{1}{2}n - 3$ bad votes, the M points from above, and minimally 0 points from the additional bribed votes):

bribed-min $:= 3 \cdot$good-min$+(\frac{1}{2}n-3) \cdot$bad-min$+M$, and this is $\frac{1}{2}nk_2 + \frac{7}{2}n + 3k_1 - 12 + M$.

Finally, our preferred candidate has exactly p-score $:= \frac{1}{4}n(\|C\| - 1) + M$ points, which is the same as $\frac{1}{4}n^2 + \frac{1}{4}n(k_1 + k_2) + M$.

The required inequality bribed-max $<$ p-score $<$ bribed-min now simplifies to $\frac{3}{4}n^2 - \frac{9}{4}n + 8 < \frac{1}{4}n^2 + (\frac{1}{4}n - 2)k_1 - \frac{1}{4}nk_2 < k_1 + \frac{7}{2}n - 12$. Substituting the definitions of k_1 and k_2, this is equivalent to $\frac{3}{4}n^2 - \frac{9}{4}n + 8 < n^3 + \frac{1}{4}n^2 < n^4 + \frac{7}{2}n - 12$, which is clearly true for large enough n. Since we can assume that the input instance has a sufficient size, the proof is completed. □

Nonapproximability of Bribery We now show that there are no efficient approximation algorithms for $-bribery-max. The following result does not only show that there is no polynomial-time approximation algorithm for the problem that achieves an approximation rate of a constant factor, it also excludes a polynomial relationship between results that can be achieved efficiently and the optimal solution.

Theorem 9 *For every polynomial q there is no polynomial-time approximation algorithm $\mathcal{A}$ for Borda-*$-bribery-*max such that for all instances I containing the election E, $\mathcal{A}$ computes a solution s such that $q(\beta(E, s)) \geq \mathrm{OPT}(I)$, unless* $\mathrm{P} = \mathrm{NP}$.

This result is significantly stronger than just excluding constant-ratio approximation algorithms: It also shows that for no constant c there is a polynomial-time approximation algorithm $\mathcal{A}$ that guarantees to produce a solution $\mathcal{A}(I)$ for every instance I containing the election E such that $\beta(E, \mathcal{A}(I))$ is at least $(\mathrm{OPT}(I))^{1/c}$. Also, the NP-hardness proven in Theorem 8 refers to an even more restricted version of the problem (where no prices are allowed), hence it does not directly follow from the non-approximability proof.

Proof. Let q be a polynomial and assume $\mathcal{A}$ is a polynomial time approximation algorithm for \$-bribery-max, such that for all instances I containing the election E: $q(\beta(E, \mathcal{A}(I))) \geq \mathrm{OPT}(I)$. We show that we can use $\mathcal{A}$ to decide 34-XC in polynomial time. Note that the construction is similar but easier than the one in the proof of Theorem 8.

Choose $d, n_0 \in \mathbb{N}$ such that $q(k) \leq k^d$ for all $k \geq n_0$. Let $S = \{s_1, \ldots, s_n\}, T_1, \ldots, T_{\frac{3}{4}n}$ be an instance of 34-XC. Without loss of generality assume that $n \geq n_0$. Let m be a natural number such that $m > n^{2d} + \frac{3}{4}n^2 - \frac{15}{4}n + 11$ and let $C = S \cup \{p, c_1, \ldots, c_m\}$ be a set of candidates, where p is our preferred candidate. Let $V = \left\{v_1, \ldots, v_{\frac{3}{4}n}\right\}$ be a set of votes with

$$v_i = p > T_i > c_1 > \cdots > c_m > S \setminus T_i$$

for every $i \in \left\{1, \ldots, \frac{3}{4}n\right\}$, and let $W = \{w_1, w_1' \ldots, w_l, w_l'\}$ be a set of votes with

$$\begin{aligned} w_i &= p > s_1 > \cdots > s_n > c_1 > \cdots > c_m, \\ w_i' &= s_n > \cdots > s_1 > p > c_m > \cdots > c_1 \end{aligned}$$

for every $i \in \{1, \ldots, l\}$. We set the price of each vote in V to 1 and the price of each vote in W to $\frac{1}{4}n+1$. The effect of W is that it leaves the relative scores of p and the candidates in S invariant, while increasing them relatively to the scores of the padding candidates $c_1, \ldots, c_m$. We introduce enough of these votes such that for every possible bribery, the candidates in S will always have more points than the padding candidates. Clearly a polynomial number of these votes suffices. The algorithm $\mathcal{A}$ cannot change these votes, since their cost exceeds the budget $\frac{1}{4}n$.

Let E be the election with candidates C and votes $V \cup W$. We apply $\mathcal{A}$ on the instance $I = (E, \frac{1}{4}n, p)$ and show that $q(\beta(E, \mathcal{A}(I))) > n^{2d}$ if and only if $S, T_1, \ldots, T_{\frac{3}{4}n}$ is a *yes*-instance of 34-XC. This shows that we can use $\mathcal{A}$ to decide the problem 34-XC, which can only happen if P = NP.

First note that $\mathrm{score}_E(p) = \frac{3}{4}n(n+m) + l(2m+n)$ and for each $s \in S$: $3(n+m-4) + l(2m+n) \leq \mathrm{score}_E(s) \leq 3(n+m-1) + (\frac{3}{4}n-3)(n-5) + l(2m+n)$.

Now let $S, T_1, \ldots, T_{\frac{3}{4}n}$ be a *yes*-instance of 34-XC and let $J \subseteq \left\{1, \ldots, \frac{3}{4}n\right\}$ specify an exact cover of S. We bribe in the following way: For every $i \in J$ we replace v_i by $v_i' = p > c_1 > \cdots > c_m > S$. Let $V' = \{v_i' \mid i \in J\} \cup \left\{v_i \mid i \in \left\{1, \ldots, \frac{3}{4}n\right\} \setminus J\right\}$ be the set of votes obtained from V with this bribe and E' the election with candidates C and votes $V' \cup W$. Since J is an exact cover, for every candidate $s \in S$ we changed exactly one of the votes in V where s was in a position among the top five candidates, therefore there are two votes left in V where s is in one of the first five positions and in all other votes in V s is voted among the last n candidates. Thus $\mathrm{score}_{E'}(s) \leq 2(n+m-1) + (\frac{3}{4}n-2)(n-1) + l(2m+n)$. Note that $\mathrm{score}_{E'}(p) = \mathrm{score}_E(p)$. It follows $q(\beta(E, A(I))) \geq \mathrm{OPT}(I) \geq m - \frac{3}{4}n^2 + \frac{15}{4}n - 11 > n^{2d}$.

Assume there is no exact cover for $S, T_1, \ldots, T_{\frac{3}{4}n}$. Note that $\mathcal{A}(I)$ changes exactly $\frac{1}{4}n$ votes from V and no vote from W. Let $E_{\mathcal{A}}$ be the election obtained from E by applying the bribing strategy $\mathcal{A}(I)$. Since there is no exact cover, there is a candidate $s \in S$ such that for all $v_i \in V$ with $s \in T_i$ it holds that v_i is not changed by $\mathcal{A}(I)$ and thus v_i is a vote in $E_{\mathcal{A}}$. That means there are at least three votes in $E_{\mathcal{A}}$ that rank s among the first 5 candidates, thus we get the lower bound $\mathrm{score}_{E_{\mathcal{A}}}(s) \geq 3(n+m-4) + l(2m+n)$. By assuming that

p is ranked first in all votes it follows $\text{score}_{E_{\mathcal{A}}}(p) \leq \frac{3}{4}n(n+m-1) + l(2m+n)$. Hence $\beta(E, \mathcal{A}(I)) \leq \frac{3}{4}n^2 - \frac{27}{4}n + 24$. W.l.o.g. we can assume that n is large enough to ensure that $\frac{3}{4}n^2 - \frac{27}{4}n + 24 \leq n^2$. Then $q(E, \beta(A(I))) \leq q(n^2) \leq n^{2d}$, concluding the proof. □

The above proof also works for a variant of the bribery problem where the voters have boolean indicators whether they can be bribed or not (instead of prices).

5 Acknowledgments

Supported in part by NSF grants CCF-0426761 and IIS-0713061, a Friedrich Wilhelm Bessel Research Award, the Alexander von Humboldt Foundation's TransCoop program, and the DAAD postdoc program. We thank the anonymous AAAI and COMSOC referees for their very helpful comments.

References

[BO91] J. Bartholdi, III and J. Orlin. Single transferable vote resists strategic voting. *Social Choice and Welfare*, 8(4):341–354, 1991.

[Bre07] E. Brelsford. Approximation and elections. Master's thesis, Rochester Institute of Technology, Rochester, NY, May 2007.

[BTT89] J. Bartholdi, III, C. Tovey, and M. Trick. The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6(3):227–241, 1989.

[BTT92] J. Bartholdi, III, C. Tovey, and M. Trick. How hard is it to control an election? *Mathematical and Computer Modeling*, 16(8/9):27–40, 1992.

[CS06] V. Conitzer and T. Sandholm. Nonexistence of voting rules that are usually hard to manipulate. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 627–634. AAAI Press, July 2006.

[CSL07] V. Conitzer, T. Sandholm, and J. Lang. When are elections with few candidates hard to manipulate? *Journal of the ACM*, 54(3):Article 14, 2007.

[DKNS01] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th International World Wide Web Conference*, pages 613–622. ACM Press, March 2001.

[DS00] J. Duggan and T. Schwartz. Strategic manipulability without resoluteness or shared beliefs: Gibbard–Satterthwaite generalized. *Social Choice and Welfare*, 17(1):85–93, 2000.

[EL05] E. Elkind and H. Lipmaa. Hybrid voting protocols and hardness of manipulation. In *The 16th Annual International Symposium on Algorithms and Computation, ISAAC 2005*, pages 206–215. Springer-Verlag *Lecture Notes in Computer Science #3872*, December 2005.

[ER97] E. Ephrati and J. Rosenschein. A heuristic technique for multi-agent planning. *Annals of Mathematics and Artificial Intelligence*, 20(1–4):13–67, 1997.

[Fal08] P. Faliszewski. Nonuniform bribery (short paper). In *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems*, pages 1569–1572, May 2008.

[FHH06] P. Faliszewski, E. Hemaspaandra, and L. Hemaspaandra. The complexity of bribery in elections. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 641–646. AAAI Press, July 2006.

[FHHR07] P. Faliszewski, E. Hemaspaandra, L. Hemaspaandra, and J. Rothe. Llull and Copeland voting broadly resist bribery and control. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, pages 724–730. AAAI Press, July 2007.

[FHS08] P. Faliszewski, E. Hemaspaandra, and H. Schnoor. Copeland voting: Ties matter. In *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems*, pages 983–990, May 2008.

[Gib73] A. Gibbard. Manipulation of voting schemes. *Econometrica*, 41(4):587–601, 1973.

[GJ79] M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1979.

[HH07] E. Hemaspaandra and L. Hemaspaandra. Dichotomy for voting systems. *Journal of Computer and System Sciences*, 73(1):73–83, 2007.

[HHR07] E. Hemaspaandra, L. Hemaspaandra, and J. Rothe. Anyone but him: The complexity of precluding an alternative. *Artificial Intelligence*, 171(5-6):255–285, April 2007.

[PR07] A. Procaccia and J. Rosenschein. Junta distributions and the average-case complexity of manipulating elections. *Journal of Artificial Intelligence Research*, 28:157–181, February 2007.

[PRZ07] A. Procaccia, J. Rosenschein, and A. Zohar. Multi-winner elections: Complexity of manipulation, control, and winner-determination. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1476–1481. AAAI Press, January 2007.

[Rus07] Nathan F. Russell. Complexity of control of borda count elections. Master's thesis, Rochester Institute of Technology, July 2007.

[Sat75] M. Satterthwaite. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10(2):187–217, 1975.

[ZPR08] M. Zuckerman, A. Procaccia, and J. Rosenschein. Algorithms for the coalitional manipulation problem. In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 277–286, January 2008.

Eric Brelsford, Edith Hemaspaandra, Henning Schnoor, Ilka Schnoor
Department of Computer Science
Rochester Institute of Technology
Rochester, NY 14623 USA Email: `eh@cs.rit.edu`

Piotr Faliszewski
Department of Computer Science
University of Rochester
Rochester, NY 14627 USA

A Deontic Logic for Socially Optimal Norms

Jan Broersen, Rosja Mastop, John-Jules Ch. Meyer, Paolo Turrini

Universiteit Utrecht
Email: paolo@cs.uu.nl

Abstract

The paper [a] discusses the interaction properties between preference and choice of coalitions in a strategic interaction. A language is presented to talk about the conflict between coalitionally optimal and socially optimal choices. Norms are seen as social constructions that enable to enforce socially desirable outcomes.

[a] This paper has been presented at the Ninth International Workshop on Deontic Logic in Computer Science (DEON'08) and it is to be found in the conference proceedings (see [4]).

1 Introduction

One fundamental issue of social choice theory [1] is how to aggregate the preferences of individual agents in order to form decisions to be taken by society as a whole. However, once we want to take into account the capabilities of agents, as we do in Multi Agent Systems, mere social choice functions are not enough to explain how and (especially) why individual interests are aggregated in the way they are. In this context, norms should be seen as social constructions that enable us to enforce socially desirable outcomes [5].

In particular there are situations in which individual preferences are not compatible and coalitions compete to achieve a given social order. A typical case is that of an agent's capability to positively or negatively affect the realization of other agents' preferences. In our paper we will view the enactment of norms as aimed at the regulation of such interactions. By enacting a norm we mean *the introduction of a normative constraint on individual and collective choices in a Multi Agent System.*

We are specifically concerned with cases where the collective perspective is at odds with the individual perspective. That is, cases where we think that letting everybody pick their own best action regardless of others' interest gives a non-optimal result. The main question we are dealing with is then: how do we determine which norms, if any, are to be imposed?

To answer this question, the paper presents a language to talk about the conflict between coalitionally optimal and socially optimal choices, and it expresses deontic notions referring to such circumstances.

1.0.1 Motivating Example

Let us consider a situation (Table 1) in which two players have the possibility of passing believed (truth) or disbelieved information (lie). If both players do not lie, they share their information, being both better off. If they both lie, they do not receive any advantage. But the worst case for a player is the one in which he does not lie and the opponent does.

In this situation, a legislator that wants to achieve the socially optimal state (players do not lie), should declare that lying is forbidden, thereby labeling the combinations of moves (lie, lie), (lie, truth) and (truth, lie) as violations.

The lying matrix is nothing but a Prisoner Dilemma [11], that is an interactive situation in which the advantages of cooperation are overruled by the incentive for individual players to defect. In Prisoner Dilemmas, individually rational players have no incentive to cooperate,

because defecting is better for a player considering all possible answers of the opponent. Note that cooperation is in the interest of the players themselves, since they would be better off than if they had pursued the unique Nash Equilibrium [11], ending up in the (lie, lie) state. However it is by no means clear that players should not pursue their own interest. In fact once we reach a state in which one player lies and the other does not lie, we cannot move to any other state without one of the two players being worse off.

1.0.2 A deontic logic for efficient interactions

Once we view a deontic language as regulating a Multi Agent System, we can say that a set of commands promote a certain interaction, prohibiting certain others. Following this line of reasoning it is possible, given a notion of optimality or efficiency, to provide a set of deontic formulas that agree with such notion, as we have done with Pareto Optimality.

What we do then is to provide a deontic language for all possible interactions, based on an underlying notion of optimality. This is quite a difference from the legal codes that we can find in a certain society, where norms are either explicitly and specifically formulated and written down in law books, contracts, etc., or are left implicit in the form of promises, values or mores [5]. The obligations and prohibitions in our system result from one general norm saying that all actions of sub-groups that do not take into account the interests of the society as a whole, are forbidden. Then, one way to use our logic is to derive obligations, permissions and prohibitions from conflicting group preferences, and use these as *suggestions* for norm introduction in the society.

We do not claim that the meaning of these operators, as studied in deontic logic, corresponds to our semantics, but rather we claim that when people make new norms they should choose those norms on the basis of the economical order behind them [17].

In order to represent abilities of agents we employ coalition logic [13], and we model an agent's preferences as a preorder on the domain of discourse. To model optimal social norms we introduce a generalization of the economical notion of Strong Pareto Efficiency (see for instance [11]), described as those sets of outcomes from which the grand coalition (i.e. the set of all agents taken together) has no interest to deviate. Our generalization consists of the fact that we do not make the assumption that these outcomes are singletons. In particular (unless specified) we do not make the assumption of playability described in [13], according to which the set of all agents can bring about any realizable outcome of the system. We consider then the elements of the complement of the efficient choices, i.e. all those that are not optimal, and we build the notion of obligation, prohibition and permission on top of them.

We postpone to future work all considerations about the effectivity of the norm, that is, all considerations about how, to what extent and in what way, the norm influences the behaviour of the agents involved.

As system designers, our aim is then to construct efficient social procedures that can guarantee a socially desirable property to be reached. We think that normative system design is at last a proper part of the Social Software enterprise [12].

Row \ Column	Truth	Lie
Truth	$(3, 3)$	$(0, 4)$
Lie	$(4, 0)$	$(1, 1)$

Table 1: Lying or not lying

The paper is structured as follows: In the first section we introduce the notions of effectivity and preference, discuss its relevant properties with respect to the problem of finding optimal social norms, and introduce the notion of domination, Pareto Efficiency and violation. In the second part we describe the syntax, the structures and the interpretation of our language. In the third part we discuss the deontic and collective ability modalities and their properties, and compare them with classical deontic and agency logics; moreover we discuss the introduction of further constraints in the models, in particular playability of the effectivity function. We show some examples to give the flavour of the situations we are able to capture with our formalism. A discussion of future work will follow and a summary of the present achievements will conclude the paper.

2 Effectivity and preference

We start by defining some concepts underlying the deontic logic of this paper. They concern the *power* and the *preferences* of collectives. We begin with the first of these, by introducing the concept of a dynamic effectivity function, adopted from [13].

2.1 Effectivity

Definition 1 (Dynamic Effectivity Function)
Given a finite set of agents Agt and a set of states W, a dynamic effectivity function *is a function* $E : W \to (2^{Agt} \to 2^{2^W})$.

Any subset of *Agt* will henceforth be called a *coalition*. For elements of W we use variables $u, v, w, \ldots$; for subsets of W we use variables $X, Y, Z, \ldots$; and for sets of subsets of W (i.e., elements of 2^{2^W}) we use variables $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \ldots$. The elements of W are called 'states' or 'worlds'; the subsets of *Agt* are called 'coalitions'; the sets of states $X \in E(w)(C)$ are called the 'choices' of coalition C in state w. The set $E(w)(C)$ is called the 'choice set' of C in w. The complement of a set $\overline{X}$ or of a choice set $\overline{\mathcal{X}}$ are calculated from the obvious domains.

A dynamic effectivity function assigns, in each world, to every coalition a set of sets of states. Intuitively, if $X \in E(w)(C)$ the coalition is said to be able to *force* or *determine* that the next state after w will be some member of the set X. If the coalition has this power, it can thus prevent that any state *not* in X will be the next state, but it might not be able to determine *which* state in X will be the next state. Possibly, some other coalition will have the power to refine the choice of C.

Many properties can be attributed to dynamic effectivity functions. An extensive discussion of them can be found in [13]. For what follows we do not need all the properties that may be considered reasonable for effectivity. However the following properties seem to be minimally required:

1. coalition monotonicity: for all X, w, C, D, if $X \in E(w)(C)$ and $C \subseteq D$, then $X \in E(w)(D)$;
2. regularity: for all X, w, C, if $X \in E(w)(C)$, then $\overline{X} \notin E(w)(\overline{C})$;
3. outcome monotonicity: for all X, Y, w, C, if $X \in E(w)(C)$ and $X \subseteq Y$, then $Y \in E(w)(C)$;
4. inability of the empty coalition: for all w, $E(w)(\emptyset) = \{W\}$

If a dynamic effectivity function has these properties, it will be called *coherent*.

The first property says that the ability of a coalition is preserved by enlarging the coalition. In this sense we do not allow new members to interfere with the preexistent

capacities of a group of agents. The second property says that if a coalition is able to force the outcome of an interaction to belong to a particular set, then no possible combinations of moves by the other agents can prevent this to happen. We think that regularity is a key property to understand the meaning of ability. If an agent is properly able to do something this means that others have no means to prevent it. The third property says that if a coalition is able to force the outcome of the interaction to belong to a particular set, then that coalition is also able to force the outcome to belong to all his supersets. Outcome monotonicity is a property of all effectivity functions in coalition logic, which is therefore a monotonic modal logic [13]. The last condition is the "Inability of the Empty Coalition". As notice also by [2] such properties forces the coalition modality for the empty coalition to be universal: intuitively the empty coalition cannot bring about non-trivial consequences.

Proposition 1 *If the effectivity function is coherent then all coalitional effectivity functions are nonempty and do not contain the empty set.*

The last property ensures that the choice of the empty coalition is always the largest possible one. This property imposes that for all $C, w, E(w)(C) \neq \emptyset$ (by coalition monotonicity) and that $\emptyset \notin E(w)(C)$ (by regularity).

2.2 Preference

Once we have defined the notion of effectivity, we also need to make reference to the preferences of coalitions. The notion of preference in strategic interaction can be understood and modeled in many ways [16]. However we believe that in strategic reasoning players need to have preferences over the possible outcomes of the game. Thus those are the preferences that constitute our main concern. Nevertheless we know from the properties of effectivity as described above, that coalitions may have different abilities at different states, in particular the grand coalition of agents may gain or lose power while changing a state: the effectivity is actually a dynamic effectivity.

The claim is thus that agents do have a fixed ordering over the domain of discourse (what we call *preferences*), and that they generate their strategic preference considering where the game may end (called *domination*). We are going to define both, discussing their properties.

We start from a preference relation for individuals over states working our way up to preferences for coalitions over sets. To do so, we start from an order on singletons, and we provide some properties to lift the relation to sets.

Definition 2 (Individual preferences for states) *A preference ordering $(\geq_i)_{i \in Agt}$ consists of a partial order (reflexive, transitive, antisymmentric) $\geq_i \subseteq W \times W$ for all agents $i \in Agt$, where $v \geq_i w$ has the intuitive reading that v is 'at least as nice' as w for agent i. The corresponding strict order is defined as usual: $v >_i w$ if, and only if, $v \geq_i w$ and not $w \geq_i v$.*

Definition 3 (Individual preferences for sets of states) *Given a preference ordering $(\geq_i)_{i \in Agt}$, we lift it to an ordering on nonempty sets of states by means of the following principles.*

1. $\{v\} \geq_i \{w\}$ *iff* $v \geq_i w$; *(Singletons)*
2. $(X \cup Y) \geq_i Z$ *iff* $X \geq_i Z$ *and* $Y \geq_i Z$; *(Left weakening)*
3. $X \geq_i (Y \cup Z)$ *iff* $X \geq_i Y$ *and* $X \geq_i Z$. *(Right weakening)*

We do not give a comprehensive specification of the logical properties of preference relations for coalitions, because this would not be relevant for the remainder of this paper. Different types of interaction may warrant the assumption of different properties for such a relation. Nevertheless, these are some properties that seem minimally required for calling some relation a preference relation. The first ensures that preferences are copied to possible choices. The properties of left and right weakening ensure a lifting from singletons to sets.

The lifting enables us to deal with preference under uncertainty or indeterminacy. The idea is that if an agent were ever confronted with two choices X, Y he would choose X over Y provided $X >_i Y$ [1].

Preferences do not consider any realizability condition, they are simply basic aspirations of individual agents, on which to construct a more realistic order on the possible outcomes of the game, which are by definition dependent on what all the agents can do together.

Out of agents' preferences, we can already define a classical notion of Pareto Efficiency.

Definition 4 (Strong Pareto efficiency) *Given a choice set $\mathcal{X}$, a choice $X \in \mathcal{X}$ is* Strongly Pareto efficient *for coalition C if, and only if, for no $Y \in \mathcal{X}$, $Y \geq_i X$ for all $i \in C$, and $Y >_i X$ for some. When $C = Agt$ we speak of* Strong Pareto Optimality*.*

We will use the characterization of Pareto Efficiency and Optimality to refer to the notions we have just defined, even though the classical definitions (compare with [11]) are weaker.

The last definition is clearer when we consider the case $\mathcal{X} = E(w)(C)$, but it is formulated in a more abstract way in order to smoothen the next two definitions.

Proposition 2 *Given the preference relation over choices $\geq_i$, and taking A, B in a choice set $\mathcal{X}$ of a coaltion C, with $PE(A)$ to indicate that the choice A is Strongly Pareto Efficient in $\mathcal{X}$, Strong Pareto Efficiency is monotonic, that is $A \subseteq B$ implies that $PE(B)$ whenever $PE(A)$.*

Proof Suppose $A \subseteq B$ and $PE(A)$ and suppose it is not the case that $PE(B)$. This means that there is a choice X in the choice set $\mathcal{X}$ of C such that $X >_i B$ for some $i \in C$. But being $A \subseteq B$ this would imply that $X >_i A$, contradicting the assumption that $PE(A)$.

Pareto Efficiency is usually defined disregarding the strategies of the players. Nevertheless, once we claim that the outcome of an interaction need not be a singleton, we need to adapt our evaluation of efficiency to such an assumption.

We now construct a preference relation on choices. To do so we first need to look at the interaction that agents' choices have with one another.

Definition 5 (Subchoice) *If E is an effectivity function, and $X \in E(w)(\overline{C})$, then the* X-subchoice set *for C in w is given by $E^X(w)(C) = \{X \cap Y \mid Y \in E(w)(C)\}$.*

As an example, let us take Table 1. Consider expressions of the form (Lie_C) to be intended as the set of worlds that make the proposition Lie_C true, with the obvious reading. In our example we have for instance the following cases:

- $E^{(Lie_C)}(w)(R) = \{(Lie_C \wedge Lie_R), (Lie_C \wedge Truth_R)\}$
- $E^{(Truth_C)}(w)(R) = \{(Truth_C \wedge Lie_R), (Truth_C \wedge Truth_R)\}$

[1] Preference lifting in interaction is also addessed by Gardenfors in [6]. A modal account of it is given in Fenrong Liu's PhD thesis [9].

Subchoices allow us to reason on a restriction of the game and to consider possible moves looking from a coalitional point of view, i.e. what is best for a coalition to do provided the others have already moved.

When agents interact therefore they make choices on the grounds of their own preferences. Nevertheless the moves at their disposal need not be all those that the grand coalition has. We can reasonably assume that preferences are filtered through a given coalitional effectivity function. That is we are going to consider what agents prefer among the things they can do.

Definition 6 (Domination) *Given an effectivity function E, X is* undominated *for C in w (abbr. $X \rhd_{C,w}$) if, and only if, (i) $X \in E(w)(C)$ and (ii) for all $Y \in E(w)(\overline{C})$, $(X \cap Y)$ is Pareto efficient in $E^Y(w)(C)$ for C.*

The idea behind the notion of domination is that if X' and X'' are both members of $E(w)(C)$ then, in principle, C will not choose X'', if X' dominates X''. This property ensures that a preference takes into account the possible moves of the other players. This resembles the notion of Individual Rationality in Nash solutions [11], according to which an action is chosen reasoning on the possible moves of the others.

Continuing our example, we have the following cases:

- $(Lie_R) \rhd_{R,w}$ for any w.
- $(Lie_C) \rhd_{C,w}$ for any w.
- not $(Lie_C, Lie_R) \rhd_{Agt,w}$

The preceding three definitions capture the idea that 'inwardly' coalitions reason Pareto-like, and 'outwardly' coalitions reason strategically, in terms of strict domination. A coalition will choose its best option given all possible moves of the opponents. Looking at the definition of Optimality we gave, we can see that undomination collapses to individual rationality when we only consider individual agents, and to Pareto efficiency when we consider the grand coalition of agents.

Proposition 3

$X \rhd_{Agt,w}$ iff X is a standard Pareto Optimal Choice in w.

$X \rhd_{i,w}$ iff X is a standard Dominating Choice in w for i.

Proof For the first, notice that since $E(w)(\emptyset) = \{W\}$ (i.e., the empty coalition has no powers), then X is undominated for Agt in w iff it is Pareto efficient in $E(w)(Agt)$ for Agt (i.e., it is Pareto optimal in w). The second is due to the restriction of undomination to singleton agents.

Nevertheless, in our framework, domination is a relation between the choice sets of a given coalition. This approach looks different from the standard one (see for instance Osborne and Rubinstein [11]) that considers instead domination as a property of states.

Proposition 4 *Game-theoretical domination is expressible in our framework.*

It is possible to rewrite a domination of a state x over a state y as the domination of the choice $\{x\}$ over $\{y\}$, making it a particular case of our definition.

Proposition 5 *Undomination is monotonic, that is for X in $E(C)(w)$ for some C, w, if $X \subseteq Y$ and $X \rhd_{C,w}$ then $Y \rhd_{C,w}$.*

which follows from monotonicity of Pareto Optimality for choices and outcome monotonicity.

2.2.1 Violation

The fundamental idea of this work is that an efficient way to impose normative constraints in a Multi Agent System is to look at the optimality of the strategic interaction of such system. In particular the presence of possible outcomes in which agents could not unanimously improve (Pareto Efficient) can be a useful guide line for designing a new set of norms to be imposed.

Following this line we define a set of violation sets as the set of those choices that are not a Pareto Efficient interaction.

Definition 7 (Violation) *If E is an effectivity function and $C \subseteq C'$, then the choice $X \in E(w)(C)$ is a violation by C towards C' in w ($X \in VIOL_{C,C',w}$) iff there is a $Y \in E(w)(C' \setminus C)$, s.t. $(X \cap Y)$ is not undominated for C' in w.*

In words, X is a violation if it is not safe for the other agents, in the sense that not all the moves at their disposal yield an efficient outcome.

We indicate with $VIOL_{C,w}$ the set violations by C at w towards Agt[2].

Proposition 6 *If C=C' then a violation is a dominated choice; If C=C'=Agt a violation is a Pareto inefficient choice.*

If we consider the Prisoner Dilemma of Table 1 the following holds:

- $(Lie_R) = VIOL_{R,w}$ for any w, since $(Lie_R \wedge Lie_C)$ is not Agt- undominated;
- $(Lie_C \wedge Lie_R) = VIOL_{Agt,w}$ for any w, since not Pareto Efficient.

We can observe here that any choice made by a single agent is a violation. The reason why it is so has to be found in the form of the game, that requires the grand coalition to form for an efficient outcome to be forced.

3 Logic

We now introduce the syntax of our logic, an extension of the language of coalition logic [13] with modalities for permission, prohibition and obligation, and a modality for rational choice.

3.1 Language

Let Agt be a finite set of agents and $Prop$ a countable set of atomic formulas. The syntax of our logic is defined as follows:

$$\phi ::= p|\neg\phi|\phi \vee \phi|[C]\phi|P(C,\phi)|F(C,\phi)|O(C,\phi)|[rational_C]\phi$$

where p ranges over $Prop$ and C ranges over the subsets of Agt. The other boolean connectives are defined as usual. The informal reading of the modalities is: "Coalition C can choose ϕ", "It is permitted (/forbidden/obligated) for coalition C to choose ϕ", "It is rational for coalition C to choose ϕ".

[2]One interesting question is whether given any dynamic effectivity function and preference relation (with the above defined properties) we can always find a coalitionally dominated action (and hence a Pareto Efficient interaction). The acquainted reader will have noticed the resemblance of this problem with that of nonemptiness of the Core [11]. We leave though to further work the analysis of this relation. In case there is none, we may consider a satisfactory notion of optimal choice - as done for instance by Horty [7] - that looks at the relation between the choices in the choice sets of each coalition.

3.2 Structures

Definition 8 (Models) *A* model *for our logic is a quadruple*

$$(W, E, \{\geq_i\}_{i\subseteq Agt}, V)$$

where:

- W *is a nonempty set of states;*
- $E : W \longrightarrow (2^{Agt} \longrightarrow 2^{2^W})$ *is a coherent effectivity function;*
- $\geq_i \subseteq W \times W$ *for each* $i \in Agt$*, is the preference relation;*
- $V : W \longrightarrow 2^{Prop}$ *is the valuation function.*

3.3 Semantics

The satisfaction relation of the formulas with respect to a pointed model M, w is defined as follows:

$$\begin{array}{rll}
M, w \models p & \text{iff} & p \in V(w) \\
M, w \models \neg\phi & \text{iff} & M, w \not\models \phi \\
M, w \models \phi \wedge \psi & \text{iff} & M, w \models \phi \text{ and } M, w \models \psi \\
M, w \models [C]\phi & \text{iff} & [[\phi]]^M \in E(w)(C) \\
M, w \models [rational_C]\phi & \text{iff} & \forall X(X \rhd_{C,w} \Rightarrow X \subseteq [[\phi]]^M) \\
M, w \models P(C, \phi) & \text{iff} & \exists X \in E(w)(C) \text{ s.t. } X \in \overline{VIOL}_{C,w} \text{ and } X \subseteq [[\phi]]^M \\
M, w \models F(C, \phi) & \text{iff} & \forall X \in E(w)(C)(X \subseteq [[\phi]]^M \Rightarrow X \in VIOL_{C,w}) \\
M, w \models O(C, \phi) & \text{iff} & \forall X \in E(w)(C)(X \in \overline{VIOL}_{C,w} \Rightarrow X \subseteq [[\phi]]^M)
\end{array}$$

In this definition, $[[\phi]]^M =_{def} \{w \in W \mid M, w \models \phi\}$.

The modality for coalitional ability is standard from Coalition Logic [13]. The modality for rational action requires for a proposition ϕ to be rational (wrt a coalition C in a given state w) that all undominated choices (for C in w) be in the extension of ϕ. This means that there is no safe choice for a coalition that does not make sure that ϕ will hold. Notice that it is still possible for a coalition to pursue a rational choice that may be socially not rational.

The deontic modalities are defined in terms of the coalitional abilities and preferences. A choice is permitted whenever it is safe, forbidden when it may be unsafe (i.e. when it contains an inefficient choice), and obligated when it is the only choice that is safe.

4 Discussion

The definition of strong permission does not allow for a permitted choice of an agent to be refined by the other agents towards a violation. In fact we define permission for ϕ as "a ϕ-choice guarantees safety from violation". A more standard diamond modality would say " doing ϕ is compatible with no violation". A "safety" definition of permission has also been studied in [18], [14], [10], [3].

4.1 Properties

It is now interesting to look at what we can say and what we cannot say within our system.

Some Validities	
1	$P(C,\phi) \rightarrow \neg O(C,\neg\phi)$
2	$F(C,\phi) \leftrightarrow \neg P(C,\phi)$
3	$P(C,\phi) \vee P(C,\psi) \rightarrow P(C,\phi \vee \psi)$
4	$O(C,\phi) \rightarrow ([C]\phi \rightarrow P(C,\phi))$
5	$[rational_C]\phi \wedge [rational_{Agt}]\neg\phi \rightarrow F(C,\phi)$
6	$O(C,\phi) \vee O(C,\psi) \rightarrow O(C,\phi \vee \psi)$
7	$O(C,\top)$
8	$F(C,\phi) \wedge F(C,\psi) \rightarrow F(C,\phi \wedge \psi)$
9	$[rational_C]\phi \wedge (\phi \rightarrow \psi) \rightarrow [rational_C]\psi$

Some non-Validities	
10	$\neg O(C,\neg\phi) \rightarrow P(C,\phi)$
11	$P(C,\phi \vee \psi) \rightarrow P(C,\phi) \vee P(C,\psi)$
12	$O(C,\phi) \leftrightarrow \neg O(C,\neg\phi)$
13	$[rational_C]\phi \leftrightarrow [rational_{Agt}]\phi$
14	$O(C,\phi) \rightarrow P(C,\phi)$
15	$O(C,\phi \vee \psi) \rightarrow O(C,\phi) \vee O(C,\psi)$

The first validity says that the presence of permission imposes the absence of contrasting obligations, but the converse is not necessarily true. The second that prohibition and permission are interdefinable. The third says that the permission of ϕ or the permission of ψ implies the permission of ϕ or ψ. The fourth that the obligation to choose ϕ for an agent plus the ability to do something entails the permission to carry out ϕ. The validity number 5 says that the presence of a safe state that is rational for the grand coalition of agents is a norm for every coalition, even in case of conflicting preferences, i.e. in case of conflict the interest of the grand coalition prevails. The sixth one that obligation for ϕ or obligation for ψ implies the obligation for ϕ or ψ. Validity 7 says that there are no empty normative systems. The next validity says that prohibition is conjunctive. The last validity says that rational moves are monotonic. This has interesting implications on the choices of the agents, since refraining, i.e. choosing the biggest possible outcome, is always rational.

It is also useful to look at the non-validities: Number 10 says that if an agent is not obliged to choose something then it is permitted to do the contrary. But of course an agent may not be able to do anything, or may be not able to refine the choices "until the optimal". The next non-validity says that a permission of choice is not equivalent to a choice of permission. Number 12 says that a coalition can be obliged to do contradictory choices. This situation happens when a coalition is powerless or optimality is not possible. The next non validity says that the rational action for a certain coalition does not necessarily coincide with that of the grand coalition. Number 14, that ought does not imply can. The last does not allow to detach specific obligation from obligatory choices.

4.1.1 Further Assumptions

Playability Our notion of agency is more general than that of game theory. In particular we assume that even the grand coalition of agents may not determine a precise outcome of the interaction.

This is due to the abandonment of the property of playability of the effectivity function, that requires, together with regularity, outcome monotonicity, coalition monotonicity that:

- $X \notin E(C)$ implies $\overline{X} \in E(\overline{C})$, that is any choice excluded to a coalition is possible for the rest of the agents (maximality);
- For all X_1, X_2, C_1, C_2 such that $C_1 \cap C_2 = \emptyset$, $X_1 \in E(C_1)$ and $X_2 \in E(C_2)$ imply that $X_1 \cap X_2 \in E(C_1 \cup C_2)$ (superadditivity)

Playability is a very strong property but it is needed to talk about games. As proved in [13] [Theorem 2.27], strategic games correspond exactly to playable effectivity functions [3]. With playable effectivity functions, the grand coalition can determine the exact outcome of the game and the dynamic effectivity function for the grand coalition of agents is the same in any state.

$$M, w \models [Agt]\phi \Leftrightarrow M \models [Agt]\phi$$

Moreover the fact that preferences do not change, induces the following stronger invariance:

$$M, w \models [rational_{Agt}]\phi \Leftrightarrow M \models [rational_{Agt}]\phi$$

So not only is every outcome reachable, but any situation shares the same social optimality. Notice that this is independent of the solution concept we may consider.

Finite Domain Another interesting assumption can be made about the finitness of the domain of discourse. With finite W, for $\mathcal{C}$ being the class of our models, we have that

Proposition 7 $\models_{\mathcal{C}} [rational_{Agt}]\phi$
implies that there exists an efficient outcome in the class of coalition models $\mathcal{C}$

Another property is the following:

$$\models_{\mathcal{C}} [rational_{Agt}]\neg\phi \wedge [C]\phi \rightarrow F(C, \phi) \tag{REG}$$

which says that any coalition has to refrain from a choice that is against an optimal state independently of its own preferences. A corresponding propery for obligation is instead the following:

$$\models_{\mathcal{C}} [rational_{Agt}]\neg\phi \wedge [C]\neg\phi \rightarrow O(C, \neg\phi) \tag{REG'}$$

[3] The proof involves the definition of strategic game as a tuple $\langle N, \{\Sigma_i | i \in N\}, o, S\rangle$ where N is a set of players, each i being endowed with a set of strategies σ_i from Σ_i, an outcome function that returns the result of playing individual strategies at each of the states in S; the definition of α effectivity function for a nonempty strategic game G, $E^{\alpha}_{G} : \wp(N) \rightarrow \wp\wp(S)$ defined as follows: $X \in E^{\alpha}_{G}\exists\sigma_C\forall\sigma_{\overline{C}}o(\sigma_C; \sigma_{\overline{C}}) \in X$. The above mentioned theorem establishes that $E^{\alpha}_{G} = E$ in case E is playable and G is a nonempty strategic game.

Coalitionally optimal norms The logic can be extended to treat norms that do not lead to a socially optimal outcome, but a coalitionally optimal outcome. That is it is possible to construct a deontic logic that pursues the interests of a particular coalition, independently of the other players' welfare. This extension is related to the work of Kooi and Tamminga on conflicting moral codes [8].

We limit the description to the obligation operator, the others are straightforward.
$M, w \models O^{C'}(C, \phi)$ iff $\forall X(X \rhd_{C,w}$ and $X \in \overline{VIOL}_{C,C',w} \Rightarrow X \subseteq [[\phi]]^M)$
where $VIOL_{C,C',w}$ is a C violation towards C', with $C \subseteq C'$.
For this operator it holds that

$$\models_{\mathcal{C}} O^C(C, \phi) \leftrightarrow [rational_C]\phi$$

that is playing for oneself boils down to rational action, and

$$\models_{\mathcal{C}} O^{Agt}(C, \phi) \leftrightarrow O(C, \phi)$$

that is, with the new operator we can express our original obligation operator.

4.2 Example: Norms of Cooperation

To consider forbidden all non optimal choices may seem a very strong requirement. Nevertheless, take the example in Table 1.

It is interesting to notice how $VIOL$ is not equivalent to the situations that each player is forbidden to choose. This is due to the fact that each player can only refine the choices of the other players, but cannot determine alone the outcome of the game: a permitted choice cannot be refined by permitted choices towards an inefficient outcome. Moreover $M \models [R]\neg(T_R) \wedge [rational_{R,C}](T_R)$, that by (REG) allows to conclude $F(R, \neg(T_R))$.

No agent is in fact obligated not to lie, but only permitted. Why is it so? Because no agent can alone reach a singleton state that is only good. But of course as a coalition $\{R, C\}$ has the obligation to end up in the optimal state.

Prisoner Dilemma, but think also of Coordination Games, in which individual players cannot reach a socially optimal outcome, have rules that say something about how coalitions should choose. This indirectly says something about the coalitions that are necessary to achieve an optimal outcome, i.e. about the coalitions that should form.

4.3 Future Work

The work here described allows for several developments. Among the most interesting ones is the study of the relation between imposed outcomes and steady states that describe where the game will actually end up (i.e. Nash Solution, the Core etc.). Conversely another feature that is worth studying is those structures in which Pareto Efficiency is not always present. Agents will reckon some actions as optimal even though there is no social equilibrium that can ever be reached. One more feature concerns the possibility of an inconsistent normative system. Further work could be done looking at the factual obedience of the norm, and how a norm affects preferences of agents (see for instance the work in [15]).

5 Conclusion

In this paper we proposed a deontic logic for optimal social norms. We described the concept of social optimality, explicitly linking it with the economical concept of Pareto Efficiency. Moreover we generalized the notion of Pareto Efficiency to capture those strategic interactions in which even the grand coalition of agents is not able to achieve every outcome.

Technically we did not assume playability of the coalitional effectivity functions. It is an important question in itself to understand the class of interactions to which such effectivity function corresponds. On top of the notion of Optimality we constructed a deontic language to talk about a normative system resulting from the imposition of such norms. We analyzed the properties of the language and discussed in details various examples from game theory and social science.

References

[1] K. Arrow. *Social Choice and Individual Values*. Yale University Press, 1970.

[2] S. Borgo. Coalitions in action logic. In *IJCAI*, pages 1822–1827, 2007.

[3] J. Broersen. Modal action logics for reasoning about reactive systems. PhD-thesis Vrije Universiteit Amsterdam, 2003.

[4] J. Broersen, J.J.Ch.Meyer, R.Mastop, and P.Turrini. A deontic logic for socially optimal norms. In *Proceedings Ninth International Workshop on Deontic Logic in Computer Science*, pages –, 2008.

[5] J. Coleman. *Foundations of Social Theory*. Belknap Harvard, 1990.

[6] P. Gardenfors. Rights, games and social choice. *Nous*, 15:341–56, 1981.

[7] J. Horty. *Deontic Logic and Agency*. Oxford University Press, 2001.

[8] B. Kooi and A.Tamminga. Conflicting obligations in multi-agent deontic logic. In J.-J. C. M. Lou Goble, editor, *Deontic Logic and Artificial Normative Systems: 8th International Workshop on Deontic Logic in Computer Science (DEON 2006)*, pages 175–186. LNCS 4048, 2006.

[9] F. Liu. *Changing for the Better: Preference Dynamics and Agent Diversity*. ILLC Dissertation Series, 2008.

[10] R. Mastop. What can you do? imperative mood in semantic theory. PhD thesis, ILLC Dissertation Series, 2005.

[11] M. Osborne and A. Rubinstein. *A course in Game Theory*. The MIT Press, 1994.

[12] R. Parikh. Social software. *Synthese*, 132(3):187–211, 2002.

[13] M. Pauly. *Logic for Social Software*. ILLC Dissertation Series, 2001.

[14] J. van Benthem. Minimal deontic logics. *Bulletin of the Section of Logic*, 8(1):36–42, 1979.

[15] J. van Benthem and F.Liu. Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logics*, 14, 2004.

[16] G. von Wright. *The logic of preference*. Edimburgh University Press, 1963.

[17] M. Weber. *Economy and Society*. Edited by Guenther Roth and Claus Wittich. New York: Bedminister Press, 1968.

[18] G. V. Wright. Deontic logic and the theory of conditions. In R. Hilpinen, editor, *Deontic Logic: Introductory and Systematic Readings*, pages 3–31, 1971.

Coalition Structures in Weighted Voting Games

Georgios Chalkiadakis and Edith Elkind and Nicholas R. Jennings

Abstract

Weighted voting games are a popular model of collaboration in multiagent systems. In such games, each agent has a weight (intuitively corresponding to resources he can contribute), and a coalition of agents wins if its total weight meets or exceeds a given threshold. Even though coalitional stability in such games is important, existing research has nonetheless only considered the stability of the grand coalition. In this paper, we introduce a model for weighted voting games with coalition structures. This is a natural extension in the context of multiagent systems, as several groups of agents may be simultaneously at work, each serving a different task. We then proceed to study stability in this context. First, we define the CS-core, a notion of the core for such settings, discuss its non-emptiness, and relate it to the traditional notion of the core in weighted voting games. We then investigate its computational properties. We show that, in contrast with the traditional setting, it is computationally hard to decide whether a game has a non-empty CS-core, or whether a given outcome is in the CS-core. However, we then provide an efficient algorithm that verifies whether an outcome is in the CS-core if all weights are small (polynomially bounded). Finally, we also suggest heuristic algorithms for checking the non-emptiness of the CS-core.

1 Introduction

Coalitional games [8] provide a rich framework for the study of cooperation both in economics and politics, and have been successfully used to model collaboration in multiagent systems [9, 3]. In such games, teams (or *coalitions*) of agents come together to achieve a common goal, and derive individual benefits from this activity.

A particularly simple, yet expressive, class of coalitional games is that of *weighted voting games (WVGs)* [11]. In a weighted voting game each player (or *agent*) has a weight, and a coalition *wins* if its members' total weight meets or exceeds a certain threshold, and loses otherwise. Weighted voting has straightforward applications in a plethora of societal and computer science settings ranging from real-life elections to computer operating systems, as well as a variety of settings involving multiagent coordination. In particular, an agent's weight can be thought of as the amount of resources available to this agent, and the threshold indicates the amount of resources necessary to achieve a task. A winning coalition then corresponds to a team of agents that can successfully complete this task.

Originally, research in weighted voting games was motivated by a desire to model decision-making in governmental bodies. In such settings, the threshold is usually at least 50% of the total weight, and the issues of interest relate to the distribution of payoffs within the *grand coalition*, i.e., the coalition of all agents. Perhaps for this reason, to date, all research on weighted voting games tacitly assumes that the grand coalition will form. However, in multiagent settings such as those described above, the threshold can be significantly smaller than 50% of the total weight, and several winning coalitions may be able to form simultaneously. Moreover, in this situation the formation of the grand coalition may not, in fact, be a desirable outcome: instead of completing several tasks, forming the grand coalition concentrates all agent resources on finishing a single task. In contrast, the overall efficiency will be higher if the agents form a *coalition structure (CS)*, i.e., a collection of several disjoint coalitions.

To model such scenarios, in this paper we introduce a model for *WVGs with coalition structures*. We then focus on the issue of *stability* in this setting. A structure is stable when rational agents are not motivated to depart from it, and thus they can concentrate on performing their task, rather than looking for ways to improve their payoffs. Therefore, stability provides a useful balance between individual goals and overall performance. To study it, we extend the notion of the *core*—a classic notion of stability for coalitional games—to our setting, by defining the *CS-core* for WVGs. We then provide a detailed study of this concept, comparing it with the classic core and analyzing its computational properties.

Our main contributions are as follows: (1) we define a new model that allows weighted voting games to admit coalition structures (Sec. 3); (2) we define the CS-core for such games, relate it to the classic core, and describe sufficient conditions for its non-emptiness (Sec. 4); (3) we show that several natural CS-core-related problems are intractable—namely, it is NP-hard to decide the non-emptiness of the CS-core and coNP-complete to check whether a given outcome is in the CS-core (Sec. 5). Interestingly, this contrasts with what holds in weighted voting games without coalition structures, where both of these problems are polynomial-time solvable; (4) we provide a polynomial-time algorithm to check if a given outcome is in the CS-core in the important special case of polynomially-bounded weights. We then show how to use this algorithm to efficiently check if a given coalition structure admits a stable payoff distribution, and suggest a heuristic algorithm to find an allocation in the core (Sec. 6). Before presenting our results, we provide some background and a brief review of related work.

2 Background and Related Work

In this section, we provide an overview of the basic concepts in coalitional game theory. Let I, $|I| = n$, be a set of players. A subset $C \subseteq I$ is called a *coalition*. A *coalitional game with transferable utility* is defined by its *characteristic function* $v : 2^I \mapsto \mathbb{R}$ that specifies the *value* $v(C)$ of each coalition C [12]. Intuitively, $v(C)$ represents the maximal payoff the members of C can jointly receive by cooperating, and it is assumed that the agents can distribute this payoff between themselves in any way.

While the characteristic function describes the payoffs available to coalitions, it does not prescribe a way of distributing these payoffs. We say that an *allocation* is a vector of payoffs $\boldsymbol{x} = (x_1, \ldots, x_n)$ assigning some payoff to each $i \in I$. We write $x(S)$ to denote $\sum_{i \in S} x_i$. An allocation is *feasible* for the grand coalition if $x(I) \leq v(I)$. An *imputation* is a feasible allocation that is also *efficient*, i.e., $x(I) = v(I)$.

A *weighted voting game (WVG)* is a coalitional game G given by a set of agents $I = \{1, \ldots, n\}$, their *weights* $\mathbf{w} = \{w_1, \ldots, w_n\}$, $w_i \in \mathbb{R}^+$, and a *threshold* $T \in \mathbb{R}$; we write $G = (I; \mathbf{w}; T)$. We use $w(S)$ to denote $\sum_{i \in S} w_i$. For a coalition $S \subseteq I$, its value $v(S)$ is 1 if $w(S) \geq T$; otherwise, $v(S) = 0$. Without loss of generality, the value of the grand coalition I is 1 (i.e., $w(I) \geq T$).

One of the best-known solution concepts describing coalitional stability is the *core*[8].

Definition 1. *An allocation $\boldsymbol{x}$ is in the* core *of G* iff $x(I) = v(I)$ **and** *for any $S \subseteq I$ we have $x(S) \geq v(S)$.*

If an allocation $\boldsymbol{x}$ is in the core, then no subgroup of agents can guarantee all of its members a higher payoff than the one they receive in the grand coalition under $\boldsymbol{x}$. This definition of the core can therefore be used to characterize the stability of the grand coalition.

The setting where several coalitions can form at the same time can be modeled using *coalition structures*. Formally, a coalition structure (CS) is an exhaustive partition of the set of agents. $\mathcal{CS}(G)$ denotes the set of all coalition structures for G. Given a structure

$CS = \{C_1, \ldots, C_k\}$, an allocation $\mathbf{x}$ is *feasible for CS* if $x(C_i) \leq v(C_i)$ for $i = 1, \ldots, k$ and *efficient for CS* if this holds with equality.

Games with coalition structures were introduced by Aumann and Dreze [2], and are obviously of interest from an AI/multiagent systems point of view, as illustrated in Section 1. Indeed, in this context dealing with coalition structures other than the grand coalition is of uttermost importance: simply put, there is a plethora of realistic application scenarios where the emergence of the grand coalition is either not guaranteed, might be perceivably harmful, or is plainly impossible. In particular, in the context of WVGs, by forming several disjoint winning coalitions, the agents generate more payoff than in the grand coalition. Additional motivation from an economics perspective is given in [2], which contains a thorough and insightful discussion on why coalition structures arise.

Now, there exists a handful of approaches in the multiagent literature that do take coalition structures explicitly into account. Sandholm and Lesser [9] discuss the stability of coalition structures when examining the problem of allocating *computational resources* to coalitions. In particular, they introduce a notion of bounded rational core that explicitly takes into account coalition structures. Apt and Radzik [1] also do not restrain themselves to problems where the outcome is the grand coalition only. Instead, they introduce various stability notions for abstract games whose outcomes can be coalition structures, and discuss simple transformations (essentially split and merge rules) by which stable partitions of the set of players may emerge. Dieckmann and Schwalbe [5] also propose a version of the core with coalition structures when dealing with coalition formation in a dynamic context. Finally, Chalkiadakis and Boutilier also define a core with coalition structures when tackling coalition formation under uncertainty [4]. None of these papers studies WVGs, however.

A thorough discussion of weighted voting games can be found in [11]. The stability-related solution concepts for WVGs (*without* coalition structures) have recently been studied by Elkind et al. [6], who also investigate them from computational perspective. However, there is no existing work in the literature studying WVGs with coalition structure—a class of games that we now proceed to define.

3 Coalition structures in WVGs

We now extend the traditional model for WVGs to allow for coalition structures. First, an *outcome* of a game is now a pair of the form (coalition structure, allocation) rather than just an allocation. Furtheremore, in the traditional model, any allocation of payoffs among the participating agents is required to be an exhaustive partition of the value of the grand coalition. In other words, it is always an imputation, i.e., an allocation of payoffs that is feasible and efficient for the grand coalition I. As we now allow WVGs to admit coalition structures, we replace the aforementioned requirement with similar requirements with respect to a coalition structure:

First, we no longer require an allocation to be an imputation in the classic sense. Instead, we demand that, for a given outcome $(CS, \boldsymbol{x})$, the allocation $\boldsymbol{x}$ of payoffs for I is feasible for CS. In this way, CS may contain *zero or more* winning coalitions. Furthermore, we define an *imputation for a coalition structure CS* as a vector $\boldsymbol{p}$ of non-negative numbers $(p_1, \ldots, p_n)$ (one for each agent in I), such that for *every* $C \in CS$ it holds $p(C) = v(C) \leq 1$; we write $\mathbf{p} \in \mathcal{I}(CS)$. That is, an imputation is now a feasible and efficient allocation of the payoff of any coalition $C \in CS$.

4 Core and CS-core of weighted voting games

In this section we define the core of WVG games with coalition structures, relate it to the "classic" core of WVG games without coalition structures, and obtain some core characterization results for a few interesting classes of WVG games.

The definition of the core (Def. 1) takes the following simple form in the traditional WVGs setting (see, e.g., [6]):

Definition 2. *The core of a WVG game $G = (I; \mathbf{w}; T)$ is the set of imputations $\boldsymbol{p}$ such that, $\forall S \subseteq I$, $w(S) \geq T \Rightarrow p(S) \geq 1$.*

Intuitively, an imputation $\boldsymbol{p}$ is in the core whenever the payoffs defined by $\boldsymbol{p}$ are such that any winning coalition already receives collective payoff of 1 (and therefore no coalition can improve its payoff by breaking away from the grand coalition).

This notion of the core cannot be directly used for coalition structures: indeed, it demands that an allocation is an imputation in the traditional sense, and therefore no imputation for a coalition structure with more than one winning coalition can ever be in the core. We will now extend this definition to the setting with coalition structures. Namely, we define the *core of weighted voting games with coalition structures*, or *CS-core*, as follows:

Definition 3. *The CS-core of a WVG game $G = (I; \mathbf{w}; T)$ with coalition structures is the set of outcomes $(CS, \boldsymbol{p})$ such that $\forall S \subseteq I$, $w(S) \geq T \Rightarrow p(S) \geq 1$* **and** *$\forall C \in CS$ it holds $p(C) = v(C)$.*

Intuitively, given an outcome that is in the CS-core, no coalition has an incentive to break away from the coalition structure.

Now, it is well-known (see, e.g., [6]) that in weighted voting games the core is non-empty if and only if there exists a *veto* player, i.e., a player that belongs to all winning coalitions, and an imputation is in the core if and only if it distributes the payoff in some way between the veto players. This directly implies the following result.

Observation 1 (An imputation in the core induces an outcome in the CS-core)**.** *Let $G = (I; \mathbf{w}; T)$. If the core of G is non-empty, then, for any $\boldsymbol{p}$ in the core, the outcome $(\{I\}, \boldsymbol{p})$ is in the CS-core of G.*

However, it turns out that the CS-core may be non-empty even when the core is empty.

Example 1. *Consider a weighted voting game $G = (I; \mathbf{w}; T)$, where $I = \{1, 2, 3\}$, $\mathbf{w} = (1, 1, 2)$ and $T = 2$. It is easy to see that none of the players in G is a veto player, so G has an empty core. On the other hand, the outcome $(CS, \mathbf{p})$, where $CS = \{\{1, 2\}, \{3\}\}$, $\mathbf{p} = (1/2, 1/2, 1)$ is in the CS-core of G. Indeed, agent 3 is getting a payoff of 1 under this outcome, so his payoff cannot improve. Therefore, the only deviation available to the other two players is to form singleton coalitions, and this is clearly not beneficial.*

We now show that if the threshold T is strictly greater than 50% the CS-core and the core coincide.

Proposition 1 (In absolute majority games, the cores coincide)**.** *Let $G = (I; \mathbf{w}; T)$ be a WVG game with $T > w(I)/2$. Then there is an outcome $(CS, \mathbf{p})$ in the CS-core of G if and only if $\mathbf{p}$ is in the core of G. Consequently, G has a non-empty core if and only if it has a non-empty CS-core.*

Proof. Suppose that an outcome $(CS, \mathbf{p})$ is in the CS-core of G. As $T > w(I)/2$, CS can contain at most one winning coalition C, and hence $p(I) = 1$. Consider any player $i \in C$ such that $p_i > 0$. If p_i is not a veto player, we have $w(I \setminus \{i\}) \geq T$, $p(I \setminus \{i\}) < 1$, so $(CS, \mathbf{p})$

is not in the CS-core of G, a contradiction. Hence, under $\mathbf{p}$ only the veto players get any payoff, which implies that $\mathbf{p}$ is in the core of G. Conversely, if $\mathbf{p}$ is in the core of G, it is easy to see that $(\{I\}, \mathbf{p})$ is in the CS-core of G. □

We can also prove the following sufficient condition for non-emptiness of the CS-core.

Theorem 1. *Any weighted voting game $G = (I; \mathbf{w}; T)$ that admits a partition of players into coalitions of weight T has a non-empty CS-core.*

Proof. Let $CS = \{C_1, \ldots, C_k\}$ be the corresponding partition such that $w(C_i) = T$ for all $i = 1, \ldots, k$. Define $\mathbf{p}$ by setting $p_j = w_j/T$ for all $j = 1, \ldots, n$. Consider any winning coalition S. We have $w(S) \geq T$, so $p(S) = w(S)/T \geq 1$, and hence S does not want to deviate. As this holds for any S with $v(S) = 1$, the outcome $(CS, \mathbf{p})$ is in the CS-core of G. □

However, it is not the case that the CS-core of a weighted voting game is always non-empty. In particular, this follows from the fact that the CS-core coincides with the core in games with $T > w(I)/2$, and such games may have an empty core. We now show that the CS-core can be empty also if $T < w(I)/2$:

Example 2. *Consider a weighted voting game $G = (I; \mathbf{w}; T)$, where $I = \{1, 2, 3, 4, 5\}$, $\mathbf{w} = (1, 1, 1, 1, 1)$ and $T = 2$. We now show that this game has empty CS-core. Indeed, consider any $CS \in \mathcal{CS}(G)$ and any $\mathbf{p} \in \mathcal{I}(CS)$. Clearly, CS can contain at most two winning coalitions, so $p(I) \leq 2$. Now, if there is a coalition $C \in CS$, $|C| \geq 3$, such that $p_i > 0$ for all $i \in C$, any two players $i, j \in C$ can deviate by forming a winning coalition and splitting the surplus $p(C \setminus \{i, j\})$. If all coalitions have size at most 2, then there is a player i that forms a singleton coalition (and hence $p_i = 0$). There also exists another player j such that $p_j < 1$ (otherwise $p(I) \geq 4$). But then $S = \{i, j\}$ satisfies $w(S) \geq T$, $p(S) < 1$, so it is a successful deviation.*

5 Non-emptiness of the CS-core: hardness results

In the rest of the paper, we deal with computational questions related to the notion of the CS-core. This topic is important since in practical applications agents have limited computational resources, and may not be able to find a stable outcome if this requires excessive computation. To provide a formal treatment of complexity issues in our setting, we assume that all weights and the threshold are integers given in binary. As any rational weights can be scaled up to integers, this can be done without loss of generality.

In the previous section, we explained how to verify whether the core is non-empty or whether a given outcome is in the core. It is not hard to see that this verification can be done in polynomial time: e.g., to check the non-emptiness of the core, we simply check if $w(I \setminus \{i\}) \geq T$ for all $i \in I$. In WVGs with coalition structures, the situation is very different. Namely, we will show that it is NP-hard to decide whether a given WVG has a non-empty CS-core. Moreover, even if we are given an imputation, it is coNP-complete to decide whether it is in the CS-core of a given WVG. We now state these computational problems more formally.

Name: NonEmptyCsCore.

Instance: Weighted voting game $G = (I; \mathbf{w}; T)$.

Question: Does G have a non-empty CS-core?

Name: InCsCore.

Instance: Weighted voting game $G = (I; \mathbf{w}; T)$, a coalition structure $CS \in \mathcal{CS}(G)$ and an imputation $\mathbf{p} \in \mathcal{I}(CS)$.

Question: Is $(CS, \mathbf{p})$ in the CS-core of G?

Both of our reductions rely on the well-known NP-complete Partition problem. An input to this problem is a pair $(A; K)$, where A is a list of positive integers $A = \{a_1, \dots, a_n\}$ such that $\sum_{i=1}^{n} a_i = 2K$. It is a "yes"-instance if there is a subset of indices J such that $\sum_{i \in J} a_i = K$ and a "no"-instance otherwise [7, p.223].

Theorem 2. *The problem* NonEmptyCsCore *is NP-hard.*

Proof. We will describe a polynomial-time procedure that maps a "yes"-instance of Partition to a "yes"-instance of NonEmptyCsCore and a "no"-instance of Partition to a "no"-instance of NonEmptyCsCore. Suppose that we are given an instance $(a_1, \dots, a_n; K)$ of Partition. If there is an i such that $a_i > K$, then obviously it is a "no"-instance of Partition, so we map it to a fixed "no"-instance of NonEmptyCsCore, e.g., by setting $G = (\{1, 2, 3, 4, 5\}; (1, 1, 1, 1, 1); 2)$ as in Example 2. Otherwise, we construct a game $G = (I; \mathbf{w}; T)$ by setting $I = \{1, \dots, n\}$, $w_i = a_i$ for $i = 1, \dots, n$, $T = K$. Note that in this case we have $w(I \setminus \{i\}) \geq T$ for any i, so there are no veto players in G.

Suppose that we have started with a "yes"-instance of Partition, and let J be such that $\sum_{i \in J} a_i = K$. Consider the coalition structure $CS = \{J, I \setminus J\}$ and an imputation $\mathbf{p}$ given by $p_i = w_i/K$ for $i = 1, \dots, n$. Note that $w(J) = w(I \setminus J) = K$, so $p(J) = p(I \setminus J) = 1$, i.e., $\mathbf{p}$ is a valid imputation. It is easy to see that $(CS, \mathbf{p})$ is in the CS-core of G. Indeed, for any winning coalition S we have $w(S) \geq K$, so $p(S) \geq 1$, i.e., the members of S would not want to deviate.

On the other hand, suppose that we have started with a "no"-instance of Partition. Consider any outcome $(CS, \mathbf{p})$ in the resulting game. Clearly, CS can contain at most one winning coalition: if there are two disjoint winning coalitions, each of them has weight K, i.e., it can be used as a "yes"-certificate for Partition. If CS contains no winning coalitions, then it is clearly unstable, as $w(I) \geq T$, $p(I) = 0$. Now, suppose that CS contains exactly one winning coalition S. In this case we have $p(S) = p(I) = 1$ and $p_i = 0$ for all $i \notin S$. We have $p_i > 0$ for some $i \in S$, so $p(I \setminus \{i\}) < 1$. Moreover, by construction, $w(I \setminus \{i\}) \geq T$. Hence, $I \setminus \{i\}$ can deviate, so $(CS, \mathbf{p})$ is not in the CS-core of G. □

Theorem 3. *The problem* InCsCore *is coNP-complete.*

Proof. We will show that the complementary problem on checking that a given outcome is not in the core is NP-complete.

First, it is easy to see that this problem is in NP: we can guess a coalition S such that $w(S) \geq T$, but $p(S) < 1$; this coalition can successfully deviate from $(CS, \mathbf{p})$.

To show that this problem is NP-hard, we construct a reduction from Partition as follows. Given an instance $(a_1, \dots, a_n; K)$ of Partition, we set $I = \{1, \dots, n, n+1, n+2\}$ and $w_i = 2a_i$ for $i = 1, \dots, n$. Define also $I' = \{1, \dots, n\}$. The weights w_{n+1} and w_{n+2} and the quota T are determined as follows. We construct a coalition S by adding agents $1, 2, \dots$ to it one by one until the weight of S is at least $2K$. If the weight of S is exactly $2K$, this means that we have started with a "yes"-instance of Partition. In this case, we set $w_{n+1} = w_{n+2} = 0$, $T = 2K$, $CS = \{I\}$, and $p_i = w_i/T$ for all $i \in I$. It is easy to see that the outcome $(CS, \mathbf{p})$ is not stable: the agents in S can deviate and increase their total payoff from $1/2$ to 1. Hence, in this case we have mapped a "yes"-instance of Partition to a "no"-instance of InCsCore.

Now, suppose that $w(S) > 2K$. As all weights are even, we have $w(S) = 2Q$ for some integer $Q > K$. Also, we have $w(I' \setminus S) = 4K - 2Q$. Set $T = 2Q$, and let $w_{n+1} = w_{n+2} =$

$2Q - 2K$. Now we have $w(I \setminus S) = 4K - 2Q + 4Q - 4K = 2Q$, i.e., both S and $I \setminus S$ are winning coalitions. Set $CS = \{S, I \setminus S\}$. Now, $\mathbf{p}$ is defined as follows: for all $i \in I'$ set $p_i = w_i/T$, set $p_{n+1} = w_{n+1}/(T+1)$, and set $p_{n+2} = 1 - p(I' \setminus S) - p_{n+1}$. We have $p(S) = w(S)/T = 1$, $p(I \setminus S) = p(I' \setminus S) + p_{n+1} + p_{n+2} = 1$, so $\mathbf{p}$ is an imputation. Note also that we have $p_{n+1} + p_{n+2} = 1 - p(I' \setminus S) = 1 - w(I' \setminus S)/T = (w_{n+1} + w_{n+2})/T$. Moreover, we have $p_{n+1} < w_{n+1}/T$, $p(I' \setminus S) = w(I' \setminus S)/T$, and hence $p_{n+2} > w_{n+2}/T$.

We now show that if $(a_1, \dots, a_n; K)$ is a "yes"-instance of Partition, then $\langle (I; \mathbf{w}; T), CS, \mathbf{p} \rangle$ is a "no"-instance of InCsCore. Indeed, suppose there is a set J such that $\sum_{i \in J} a_i = K$. Consider the coalition $J' = J \cup \{n+1\}$. We have $w(J') = 2K + 2Q - 2K$, so it is a winning coalition. On the other hand, $p(J') = p(J) + p_{n+1} = w(J)/T + w_{n+1}/(T+1) < w(J')/T = 1$. Hence, J' can benefit from deviating, i.e., $(CS, \mathbf{p})$ is not in the core.

On the other hand, suppose that $\langle (I; \mathbf{w}; T), CS, \mathbf{p} \rangle$ is a "no"-instance of InCsCore, i.e., there is a set J'' such that $w(J'') \geq T$, $p(J'') < 1$. Suppose that $w(J'') > T$, i.e., $w(J'') \geq T + 1$. We have $p_i \geq w_i/(T+1)$ for all $i \in I$ (indeed, we have $p_i \geq w_i/T$ for $i \neq n+1$ and $p_i = w_i/(T+1)$ for $i = n+1$), so $p(J'') \geq w(J'')/(T+1) \geq 1$, a contradiction. Hence, we have $w(J'') = T$. Moreover, if $n+1 \notin J''$, we have $p(J'') \geq w(J'')/T = 1$, a contradiction again. Therefore, $n+1 \in J''$. Finally, if $n+2 \in J''$, we have $p(J'') = p(J'' \cap I') + p_{n+1} + p_{n+2} = w(J'' \cap I')/T + (w_{n+1} + w_{n+2})/T = w(J'')/T = 1$, also a contradiction. We conclude that $w(J'') = T$, $n+1 \in J''$, $n+2 \notin J''$, and hence $w(J'' \cap I') = 2Q - (2K - 2Q) = 2K$, which means that $\sum_{i \in J'' \cap I'} a_i = K$, i.e., $J'' \cap I'$ is a witness that we have a "yes"-instance of Partition. □

6 Algorithms for the CS-core

The hardness results presented in the previous section rely on all weights being given in binary. However, in practical applications it is often the case that the weights are not too large, or can be rounded down so that the weights of all agents are drawn from a small range of values. In such cases, we can assume that the weights are given in unary, or, alternatively, are at most polynomial in n. It is therefore natural to ask if our problems can be solved efficiently in such settings. It turns out that for InCsCore this is indeed the case.

Theorem 4. *There exists a pseudopolynomial*[1] *algorithm* $\mathcal{A}_{\text{InCsCore}}$ *for* InCsCore*, i.e., an algorithm that correctly decides whether a given outcome* $(CS, \mathbf{p})$ *is in the CS-core of a weighted voting game* $(I; \mathbf{w}; T)$ *and runs in time* $\text{poly}(n, w(I), |\mathbf{p}|)$*, where* $|\mathbf{p}|$ *is the number of bits in the binary representation of* $\mathbf{p}$*.*

Proof. The input to our algorithm is an instance of InCsCore, i.e., a weighted voting game $G = (I; \mathbf{w}; T)$, a coalition structure $CS \in \mathcal{CS}(G)$ and an imputation $\mathbf{p} \in \mathcal{I}(CS)$. The outcome $(CS, \mathbf{p})$ is not stable if and only if there exists a set S such that $w(S) \geq T$, but $p(S) < 1$. This means that our problem is essentially reducible to the classic Knapsack problem [7], which is known to have a pseudopolynomial time algorithm based on dynamic programming. In what follows, we present this algorithm for completeness.

Let $W = w(I)$. For $j = 1, \dots, n$ and $w = 1, \dots, W$, let $P(j, w)$ be the smallest total payoff of a coalition with total weight w all of whose members appear in $\{1, \dots j\}$: $P(j, w) = \min\{p(J) \mid J \subseteq \{1, \dots, j\}, w(J) = w\}$. Now, if $\min_{w=T,\dots,W} P(n, w) < 1$, it means that there is a winning coalition whose total payoff is less than 1. Obviously, this coalition would like to deviate from $(CS, \mathbf{p})$, i.e., in this case $(CS, \mathbf{p})$ is not in the CS-core. Otherwise, the payoff to any winning coalition (not necessarily in CS) is at least 1, so no group of agents wants to deviate from CS, and thus $(CS, \mathbf{p})$ is in the CS-core.

[1] An algorithm whose running time is polynomial if all numbers in the input are given in unary is called *pseudopolynomial*.

It remains to show how to compute $P(j, w)$ for all $j = 1, \ldots, n$, $w = 1, \ldots, W$. For $j = 1$, we have $P(1, w) = p_1$ if $w = w_1$ and $P(1, w) = +\infty$ otherwise. Now, suppose we have computed $P(j, w)$ for all $w = 1, \ldots, W$. Then we can compute $P(j + 1, w)$ as $\min\{P(j, w), p_{j+1} + P(j, w - w_j)\}$. The running time of this algorithm is polynomial in n, W and $|\mathbf{p}|$, i.e., in the size of the input. □

We now show how to use the algorithm $\mathcal{A}_{\text{InCsCore}}$ to check whether for a given coalition structure CS there *exists* an imputation $\mathbf{p}$ such that the outcome $(CS, \mathbf{p})$ is in the CS-core. Our algorithm for this problem also runs in pseudopolynomial time.

Theorem 5. *There exists a pseudopolynomial algorithm $\mathcal{A}_p$ that given a weighted voting game $G = (I; \mathbf{w}; T)$ and a coalition structure $CS \in \mathcal{CS}(G)$, correctly decides whether there exists an imputation $\mathbf{p} \in \mathcal{I}(CS)$ such that the outcome $(CS, \mathbf{p})$ is in the CS-core of G and runs in time* $\text{poly}(n, w(I))$.

Proof. Suppose $CS = \{C_1, \ldots, C_k\}$. Consider the following linear feasibility program (LFP) with variables $p_1, \ldots, p_n$:

$$\begin{aligned}
p_i &\geq 0 && \text{for all } i = 1, \ldots, n \\
\sum_{i \in C_j} p_i &= 1 && \text{for all } j \text{ such that } w(C_j) \geq T \\
\sum_{i \in C_j} p_i &= 0 && \text{for all } j \text{ such that } w(C_j) < T \\
\sum_{i \in J} p_i &\geq 1 && \text{for all } J \subseteq I \text{ such that } w(J) \geq T
\end{aligned} \tag{1}$$

The first three groups of equations require that $\mathbf{p}$ is an imputation for CS: all payments are non-negative, the sum of payments to members of each winning coalition in CS is 1, and the sum of payments to members of each losing coalition in CS is 0. The last group of equations states that there is no profitable deviation: the payoff to each winning coalition (not necessarily in CS) is at least 1. Clearly, we can implement the algorithm $\mathcal{A}_p$ by solving this LFP, as follows:

The size of this LFP may be exponential in n, as there is a constraint for each winning coalition. Nevertheless, it is well-known that such LFPs can be solved in polynomial time by the ellipsoid method provided that they have a polynomial-time *separation oracle*. A separation oracle is an algorithm that, given an alleged feasible solution, checks whether it is indeed feasible, and if not, outputs a violated constraint [10]. In our case, such an oracle will have to verify whether a given vector $\mathbf{p}$ violates one of the constraints in (1):

It is straightforward to verify whether all p_i are non-negative, and whether the payment to each winning coalition in CS is 1 and the payment to each losing coalition in CS is 0. If any of these constraints is violated, our separation oracle outputs the violated constraint. If this is not the case, we can use the algorithm $\mathcal{A}_{\text{InCsCore}}$ described in the proof of Theorem 4 to decide whether there exists a winning coalition J such that $w(J) \geq T$, $p(J) < 1$; this algorithm can be easily adapted to return such coalition if one exists. If $\mathcal{A}_{\text{InCsCore}}$ produces such a coalition, our separation oracle outputs the corresponding violated constraint. If $\mathcal{A}_{\text{InCsCore}}$ reports that no such coalition exists, then $(CS, \mathbf{p})$ is in the CS-core of G, so we can output $\mathbf{p}$ and stop. □

The algorithm $\mathcal{A}_p$ described in the proof of Theorem 5 allows us to check whether a given weighted voting game G has a non-empty CS-core: we can enumerate all coalitional structures in $\mathcal{CS}(G)$, and for each of them check whether there is an imputation $\mathbf{p}$, which, combined with the coalition structure under consideration, results in a stable outcome.

However, the number of coalition structures in $\mathcal{CS}(G)$ is exponential in n, and solving a linear feasibility problem for each of them using the ellipsoid method is prohibitively expensive. We now describe heuristics that can be used to speed up this process.

First, observe that we can exclude from consideration coalition structures that contain more than one losing coalition. Indeed, if any such coalition structure is stable, the coalition structure obtained from it by merging all losing coalitions will also be stable. Moreover, we can assume that each winning coalition C in our coalition structure is *minimal*, i.e., if we delete any element from C, it becomes a losing coalition. The argument is similar to the previous case: if any coalition structure with a non-minimal coalition C is stable, the coalition structure obtained by moving the extraneous element from C to the (unique) losing coalition is also stable.

Now, suppose that we have a coalition structure $CS = \{C_0, C_1, \ldots, C_k\}$ such that $v(C_0) = 0$ (C_0 can be empty), $v(C_i) = 1$ for $i = 1, \ldots, k$, and all C_i, $i > 0$, are minimal. Consider an agent $j \in C_i$, $i > 0$. If $p_j > 0$ and $w(C_0) \geq w_j$, then CS is not stable: the players in $C_0 \cup C_i \setminus \{j\}$ can deviate by forming a winning coalition and redistributing the extra payoff of p_j between themselves. Set $C_i' = \{j \in C_i \mid w_j \leq w(C_0)\}$. The argument above shows that the members of the sets C_i' get paid 0 under any imputation $\mathbf{p}$ such that $(CS, \mathbf{p})$ is stable. Now, set $C' = \cup_{i>0} C_i'$. If $w(C') + w(C_0) \geq T$, there is no imputation $\mathbf{p}$ such that $(CS, \mathbf{p})$ is stable: any such imputation would have to pay 0 to players in C_0 and each C_i', but then the players in these sets can jointly deviate and form a winning coalition.

Therefore, we can speed up the algorithm in the proof of Theorem 5 as follows: given a coalition structure $CS = \{C_0, C_1, \ldots, C_k\}$, compute the sets C_i', $i = 1, \ldots, k$, and check whether $w(C') + w(C_0) \geq T$. If this is indeed the case, there is no imputation $\mathbf{p}$ such that $(CS, \mathbf{p})$ is stable. Otherwise, run the algorithm $\mathcal{A}_p$. Clearly, this preprocessing step is very fast (in particular, unlike $\mathcal{A}_p$, it runs in polynomial time even if the weights are large, i.e., given in binary), and in many cases we will be able to reject a candidate coalition structure without having to solve the LFP (which is computationally expensive).

7 Conclusions

In this paper, we extended the model of weighted voting games (WVGs) to allow for the formation of coalition structures, thus permitting more than one coalition to be *winning* at the same time. We then studied the problem of stability of the resulting structure in such games. Specifically, we introduced *CS-core* (the core with coalition structures), and discussed its properties by relating it to the traditional concept of the core for WVGs and proving sufficient conditions for its non-emptiness. Following that, we showed that deciding CS-core non-emptiness or checking whether an outcome is in the CS-core are computationally hard problems (unlike what holds in the traditional WVGs setting). However, for specific classes of games, we presented polynomial-time algorithms for checking if a given outcome is in the CS-core, and discovering a CS-core element given a coalition structure. We then suggested heuristics that, combined with these algorithms, can be used to generate an outcome in the CS-core. We believe that the line of work presented here is important: Weighted voting games are well understood, and the addition of coalition structures increases the usability of this intuitive framework in multiagent settings (where weights can represent resources and thresholds do not necessarily exceed 50%).

In terms of future work, we intend, first of all, to come up with new heuristics to speed up our algorithms. In addition, notice that the algorithms and heuristics of Sec. 6 provide essentially centralized solutions to their respective problems. Therefore, we are interested in studying *decentralized* approaches; to begin, we intend to speed up, in the WVGs context, the exponential decentralized coalition formation algorithm of [5]. Finally, studying other

solution concepts in this context, such as the Shapley value [8], is also within our intentions.

Acknowledgements This research was undertaken as part of the ALADDIN (Autonomous Learning Agents for Decentralised Data and Information Networks) project. ALADDIN is jointly funded by a BAE Systems and EPSRC strategic partnership (EP/C548051/1).

References

[1] K. Apt and T. Radzik. 'Stable Partitions in Coalitional Games', working paper, available at http://arxiv.org/abs/cs.GT/0605132, 2006.

[2] R.J. Aumann and J.H. Dreze, 'Cooperative Games with Coalition Structures', *International Journal of Game Theory*, **3**(4), 217–237, 1974.

[3] P. Caillou, S. Aknine, and S. Pinson, 'Multi-agent Models for Searching Pareto Optimal Solutions to the Problem of Forming and Dynamic Restructuring of Coalitions', in *Proc. of ECAI'02*.

[4] G. Chalkiadakis and C. Boutilier, 'Bayesian Reinforcement Learning for Coalition Formation Under Uncertainty', in *Proc. of AAMAS'04*.

[5] T. Dieckmann and U. Schwalbe. Dynamic Coalition Formation and the Core, Economics Department Working Paper Series, Department of Economics, National University of Ireland - Maynooth, 1998.

[6] E. Elkind, L.A. Goldberg, P.W. Goldberg, and M. Wooldridge, 'Computational Complexity of Weighted Threshold Games', in *Proc. of AAAI'07*.

[7] M. Garey and D. Johnson, *Computers and Intractability; A Guide to the Theory of NP-Completeness*, W. H. Freeman & Co., N. York, 1990.

[8] R.B. Myerson, *Game Theory: Analysis of Conflict*, Harvard University Press, 1991.

[9] T. Sandholm and V.R. Lesser, 'Coalitions Among Computationally Bounded Agents', *Artificial Intelligence*, **94**(1), 99 – 137, 1997.

[10] A. Schrijver, *Combinatorial Optimization: Polyhedra and Efficiency*, Springer, 2003.

[11] A. Taylor and W. Zwicker, *Simple Games: Desirability Relations, Trading, Pseudoweightings*, Princeton University Press, Princeton, 1999.

[12] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, 1944.

Georgios Chalkiadakis
School of Electronics and Computer Science
University of Southampton
Southampton, United Kingdom
Email: gc2@ecs.soton.ac.uk

Edith Elkind
School of Electronics and Computer Science
University of Southampton
Southampton, United Kingdom
Email: ee@ecs.soton.ac.uk

Nicholas R. Jennings
School of Electronics and Computer Science
University of Southampton
Southampton, United Kingdom
Email: nrj@ecs.soton.ac.uk

Compiling the votes of a subelectorate

Yann Chevaleyre, Jérôme Lang, Nicolas Maudet & Guillaume Ravilly-Abadie

Abstract

In many practical contexts where a number agents have so as to find a common decision, the votes do not come all together at the same time (for instance, when voting about a date for a meeting, it often happens that one or two participants express their preferences later than others). In such situations, we might want to preprocess the information given by the subelectorate (consisting of those voters who have expressed their votes) so as to "compile" the known votes for the time when the latecomers will have expressed their votes. We study the amount of space necessary to such a compilation, in function of the voting rule used, the number of candidates, the number of voters who have already expressed their votes and the number of remaining voters. We position our results with respect to existing work, especially on vote elicitation and communication complexity.

1 Introduction

In many practical contexts where a number agents have so as to find a common decision, the votes do not come all together at the same time. For instance, in some political elections, the votes of the citizens living abroad is known only a few days after the rest of the votes. Or, when voting about a date for a meeting, it often happens that one or two participants express their preferences later than the others. In such situations, we might want to preprocess the information given by the subelectorate (consisting of those voters who have expressed their votes) so as to prepare the ground for the time when the latecomers will have expressed their votes. What does "preparing the ground" exactly mean? We may think of two different criteria:

- *space*: synthesize the information contained in the votes of the subelectorate, using *as less space as possible*, while keeping enough information so as to be able to compute the outcome once the newcomers hace expressed their votes;
- *on-line time*: compile the information, using as much off-line time and space as needed, in such a way that once the newcomers hace expressed their vote, the outcome can be computed *as fast as possible*.

These two criteria not only differ, but are, to some extent, opposed.

The research area of *knowledge compilation* (see for instance [3, 6]) lay the focus on on-line space and typically looks for worst-case exponentially large rewritings of the "fixed part" of the input, enabling on-line time complexity to fall down. While knowledge compilation is definitely relevant to voting (the fixed part being the known votes, and the varying part the votes of the latecomers), and would surely deserve a paper on its own, in this paper, however, we focus on minimizing space (and do not care about on-line time).

While should we care about synthesizing the votes of a subelectorate in as less space as possible? After all, one may think, the current cost of storage is so low that one should not care about storing millions of votes. There are two possible objections to this line of argumentation. The first one has to do with the size of the candidate set. In one-seat political elections, the number of candidates is typically no more than a dozen; however, in "profane" votes, such as multiple elections [2], the set of candidates has a combinatorial structure and can be extremely large (possibly much more than a few millions – while it is

difficult to imagine an election with more than a few million voters). The second objection has to do with the practical acceptance of the voting rule. Suppose the electorate is split into different districts (generally, corresponding to geographical entities). Each district can count its ballots separately and communicate the partial outcome to the central authority (e.g. the Ministry of Innner Affairs), which, after gathering the outcomes from all districts, will determine the final outcome. The space needed to synthesize the votes of a district (with respect to a given voting rule) is precisely the amount of information that the district has to send to the central authority. Now, it is important that the voters should be able to check as easily as possible the outcome of the election. Take a simple rule, such as plurality or Borda. Obviously, it is enough (and almost necessary, as we see later) for each district to send only its "local" plurality or Borda scores to the central authority. If the district is small enough, it is not difficult for the voters of this district to check that the local results are sound (for instance, each political party may delegate someone for checking the ballots); provided these local results are made public (which is usually the case – in most countries, they are published in newspapers), every voter can check the final outcome from these local outcomes (in the case of plurality or Borda, simply by summing up the local scores). Clearly, if the information about the votes of a district being necessary for computing the final outcome is large (*e.g.*, if one needs to know how many voters have expressed every possible linear order on the candidate set), it will be impractical to publish the results locally, and therefore, difficult to check the final outcome, and voters may then be reluctant to accept the voting rule. Although the compilation of the votes of a subelectorate has not been considered before (as far as we know), several related problems have been investigated:

- the *complexity of vote elicitation* [4]: given a voting rule r, a set of known votes S, and a set of t new voters, is the outcome of the vote already determined from S?
- the *computation of possible and necessary winners* [7, 11, 9, 10]: given a voting rule r, a set of incomplete votes (that is, partial orders on the set of candidates), who are the candidates who can still possibly win the election, and is there a candidate who surely wins it?
- the *communication complexity of voting rules* [5]: given a voting rule r and a set of voters, what is the worst-case cost (measured in terms of number of bits transmitted) of the best protocol allowing to compute the outcome of the election?

In the first two cases, the connection is clear. In the extremely favourable case where the outcome of the vote is already determined from S (corresponding to the existence of a necessary winner, or to a positive answer to the vote elicitation problem), the space needed to synthesize the input is just the binary encoding of the winner. The connection with communication complexity [8] will be discussed more explicitly in Section 2, after the notion of compilation is introduced formally. Then in Section 3 we determine the compilatioon complexity of some of the most common voting rules.

2 Compilation complexity as one-round communication complexity

Let X be a finite set of *candidates* and N a finite set of *voters*. Let $p = |X|$ and $n = |N|$. A *vote* is a linear order over X. We sometimes denote votes in the following way: $a \succ b \succ c$ is denoted by abc, etc. For $m \leq n$, a (p, m)-*profile* is a tuple $P = \langle V_1, \ldots, V_m \rangle$ where each V_i is a vote. When $m < n$ (resp. $m = n$), we call such profiles *partial* (resp. *complete*). Let $\mathcal{P}_X^m$ be the set of all m-voters profiles over X. A voting rule is a function r from $\mathcal{P}_X^n$ to X.

As the usual definition of most common voting rules does not exclude the possibility of ties, we assume these ties are broken by a fixed priority order on candidates.

We now consider situations where only some of the voters (the "subelectorate") have expressed their votes. Let $m \leq n$ number of voters who have expressed their vote, and $P \in \mathcal{P}_X^m$ the partial profile obtained from these m voters. We say that two partial profiles are r-equivalent if no matter the remaining unknown votes, they will lead to the same outcome. We distinguish between two cases, depending on whether the number of remaining voters is fixed or not.

Definition 1 *Let $P, Q \in \mathcal{P}_X^m$ be two m-voters X-profiles and r a voting rule. We say that*

- *given $k \geq 0$, P and Q are* (r,k)-equivalent *if for every $R \in \mathcal{P}_X^k$ we have $r(P \cup R) = r(Q \cup R)$.*
- *P and Q are* r-equivalent *if they are (r,k)-equivalent for every $k \geq 0$.*

Example 1 *Let r_P be the plurality rule and r_B the Borda rule, $X = \{a,b,c\}$ and $m = 4$. Let $P_1 = \langle abc, abc, abc, abc\rangle$, $P_2 = \langle abc, abc, acb, acb\rangle$, $P_3 = \langle acb, acb, abc, abc\rangle$ and $P_4 = \langle abc, abc, abc, bca\rangle$. Then we have the following:*

- *P_2 and P_3 are r_P-equivalent and r_B-equivalent. More generally, they are r-equivalent for every anonymous voting rule r.*
- *P_1 and P_2 are r_P equivalent. They are also (r_B,k)-equivalent for every $k \leq 2$. However they are not (r_B,k)-equivalent for $k \geq 2$. For $k = 3$, this can be seen by considering $R = \langle bca, bca, bca\rangle$. We have $r_B(P_1 \cup R) = b$ but $r_B(P_2 \cup R) = a$; therefore, P_1 and P_2 are not r_B-equivalent.*
- *P_1 and P_4 are (r_P,k)-equivalent for every $k \leq 2$, but not for $k \geq 2$, therefore they are not r_P-equivalent (nor r_B-equivalent).*

We denote (r,k)-equivalence and r-equivalence by, respectively, $\sim_{r,k}$ and $\sim_r$. Obviously, $\sim_{r,k}$ and $\sim_r$ are transitive, therefore they are indeed equivalence relations. We now define the *compilation complexity* of a voting rule. We have two notions, depending on whether the number of remaining candidates (i.e. the size of R) is fixed or not.

Definition 2 *Given a voting rule r, we say that a function σ from $\mathcal{P}_X^m$ to $\{0,1\}^*$ is a* compilation function *for (r,k) if there exists a function $\rho : \{0,1\}^* \times \mathcal{P}_X^k \to X$ such that for every $P \in \mathcal{P}_X^m$ and every $R \in \mathcal{P}_X^k$, $\rho(\sigma(P), R) = r(P \cup R)$. The size of σ is defined by $Size(\sigma) = \max\{\sigma(P) \mid P \in \mathcal{P}_X^m\}$. The* compilation complexity *of (r,k) is then defined by*

$$C(r,k) = \min\{Size(\sigma) \mid \sigma \text{ is a compilation function for } (r,k)\}$$

Informally, the compilation complexity of (r,k) is the minimum space needed to compile the m-voter partial profile P without knowing the remaining k-voter profile R. This notion does not take into account the off-line time needed to compute σ, nor the off-line time needed to compute ρ. The usual knowledge compilation view would focus on minimizing the time needed to compute ρ,regardless of the size of σ (and the time needed to compute it). The definitions when k is not fixed are similar:

Definition 3 *Given a voting rule r, we say that a function σ from $\mathcal{P}_X^m$ to $\{0,1\}^*$ is a* compilation function *for r if there exists a function $\rho : \{0,1\}^* \times \mathcal{P}_X^* \to X$, where $\mathcal{P}_X^* = \cup_{k \geq 0} \mathcal{P}_X^k$, such that for every $P \in \mathcal{P}_X^m$, every $k \geq 0$ and every $R \in \mathcal{P}_X^k$, $\rho(\sigma(P), R) = r(P \cup R)$. The* compilation complexity *of r is defined by*

$$C(r) = \min\{Size(\sigma) \mid \sigma \text{ is a compilation function for } r\}$$

An equivalent way of seeing compilation complexity is related to multiparty communication complexity. When n agents have to compute a function f, while each of them only knows a part of the input, the deterministic communication complexity (see [8]) of f is the worst-case number of bits that the agents have to exchange so as to be able to know the outcome. The communication complexity of common voting rules is identified in [5].

While standard communication complexity does not impose any restriction on the protocol that the agents may use to compute f, imposing such restrictions leads to variants of communication complexity; especially, a *one-round protocol* for two agents A and B is a protocol where A sends only one message to B, and then B sends the output to A (see Section 4.2 of [8]). The *one-round communication complexity* of f is the worst-case number of bits of the best one-round protocol for f. This is exactly the same as the compilation complexity of f, up to a minor difference: we do not care about B sending back the output to A. Here, A represents the set of voters having already expressed their votes, and B the remaining voters; the space needed to synthesize the votes of A is the amount of information that A must send to B so that B can be able to compute the final outcome[1].

We have this following general characterization of compilation complexity. Up to minor details, this is a reformulation of Exercise 4.18 in [8]. For the sake of the exposition, we reformulate it in our own terms and include its proof.

Proposition 1 *Let r be a voting rule. Let m be the number of initial voters and p the number of candidates.*

- *given $k \geq 0$, if the number of equivalence classes for $\sim_{r,k}$ is $f(m,p,k)$ then the compilation complexity of (r,k) is exactly $\lceil \log f(m,p,k) \rceil$.*
- *if the number of equivalence classes for the r-equivalence relation $\sim_r$ is $g(m,p)$ then the compilation complexity of r is exactly $\lceil \log g(m,p) \rceil$.*

Proof: We give the proof only for the case of (r,k); the proof with unbounded k is similar. We first show that $C(r,k) \geq \lceil \log f(m,p,k) \rceil$. Suppose $\sim_{k,r}$ has $f(m,p,k)$ equivalence classes. Assume there is a number $\theta < \lceil \log f(m,p,k) \rceil$, a function $\sigma : \mathcal{P}_X^m \to \{0,1\}^\theta$ and a mapping $\rho : \{0,1\}^* \to X$ such that for every $P \in \mathcal{P}_X^m$ and $R \in \mathcal{P}_X^k$, $\rho(\sigma(P),R) = r(P \cup R)$. We first note that $\theta < \lceil \log f(m,p,k) \rceil$ implies $\theta < \log f(m,p,k)$. Let $\approx_\sigma$ be the equivalence relation on $\mathcal{P}_X^m$ defined by $P \approx_\sigma Q$ if $\sigma(P) = \sigma(Q)$. Because for every P, $|\sigma(P)| \leq \theta$, $\approx_\sigma$ has at most 2^θ equivalence classes. Since $2^\theta < f(m,p,k)$, $\approx_\sigma$ has strictly less equivalence classes than $\sim_{k,r}$. Hence there exists a pair (P,Q) such that $\sigma(P) = \sigma(Q)$ but $P \not\sim_{k,r} Q$. $P \not\sim_{k,r} Q$ means that there exists a profile $R \in \mathcal{P}_X^k$ such that $r(P \cup R) \neq r(Q \cup R)$. Now, $r(P \cup R) = \rho(\sigma(P),R) = \rho(\sigma(Q),R) = r(Q \cup R)$, hence a contradiction. We now show that $C(r,k) \leq \lceil \log f(m,p,k) \rceil$. Let us enumerate and number all $f(m,p,k)$ equivalence classes for $\sim_{k,r}$. For every P, let $i(P)$ be the index of its equivalence class for $\sim_{k,r}$. Define the translation $\sigma(P) = i(P)$. We note that the size of σ is exactly $\lceil \log f(m,p,k) \rceil$. Now, define

[1]Since one-round communication complexity is never smaller than standard communication complexity, we expect the lower communication complexity bounds communication in [5] to be lower bounds of compilation complexity. However, making this more precise is not so simple, because in [5] there is no partition between two subelectorates: their results mention only the total number of candidates, whereas ours mention the number of candidates who have already expressed their votes. Let $D(r,n,p)$ the (deterministic) communication complexity of r for n voters and p candidates as in [5]. Let us now introduce this variant of communication complexity: if $m \leq n$, define $D(r,n,m,p)$ as the cost of the optimal protocol for computing r, where only the bits sent by the m first voters count for the cost of a protocol (the remaining $n-m$ can communicate for free). Obviously, we have $D(r,n,m,p) \leq D(r,n,p)$. Moreover, if $C(r,m,p)$ is the compilation complexity of r for m voters and p candidates then for every $n \geq m$ we have $C(r,m,p) \geq D(r,n,m,p)$. In order to conclude $C(r,m,p) \geq D(r,m,p)$, we would have to show that for all voting rules considered here, we have $D(r,n,m,p) = D(r,m,p)$ (which we conjecture).

ρ by $\rho(j, R) = r(P \cup R)$ for an arbitrary P such that $i(P) = j$. The result follows. ■

Here are now a few simple results about voting rules in general.

Proposition 2 *Let r be a voting rule, and r' an anonymous voting rule.*

- $C(r) \leq m \log(p!)$;
- $C(r') \leq \min(m \log(p!), p! \log m)$.

The proof is easy. For any r, the number of equivalence classes cannot be larger than the number of profiles, and there are $(p!)^m$ possible profiles. For any anonymous r, the bound $p! \log m$ comes from the fact that linear orders on X can be enumerated, together with the number of voters who choose it. $p! \log m$ can be smaller than $m. \log(p!)$ when m becomes large enough and p small enough. At the other extremity of the spectrum, we have:

Proposition 3

- *the compilation complexity of a dictatorship is* $\log p$;
- *the compilation complexity of r is* 0 *if and only if r is constant.*

In these limit cases, whether we know or not the number of remaining voters is irrelevant.

3 Some case studies

We now consider a few specific families of voting rules. For each of these we adopt the following methodology: we first seek a characterization of the equivalence classes for the given rule, then we use this characterization to count the number of equivalence classes. In simple cases, it will be easy to enumerate exactly these classes and Proposition 1 will give us the exact compilation complexity of the rule. In more complex cases, we will exhibit a simple upper bound and provide a lower bound of the same order.

3.1 Plurality and Borda

Let $\vec{s} = \langle s_1, \ldots, s_n \rangle$ be a vector of integers such that $s_1 \geq s_2 \geq \ldots \geq s_n = 0$. The scoring rule induced by $\vec{s}$ is defined by: for every candidate x, $score_{\vec{s}}(x, P) = \sum_{i=1}^{n} s_i.n(P, i, x)$, where $n(P, i, x)$ is the number of votes in P that rank x in position i; and $r_{\vec{s}}(P)$ is the candidate maximizing $score_{\vec{s}}(x, P)$ (in case of a tie, a priority relation on candidates is applied). The plurality (resp. Borda) rule r_P (resp. r_B) is the scoring rule corresponding to the vector $\langle 1, 0, \ldots, 0 \rangle$ (resp. $\langle p-1, p-2, \ldots, 0 \rangle$).

Plurality. We begin with the compilation complexity of plurality (antiplurality is similar).

Lemma 1 *For $P \in \mathcal{P}_X^m$ and $x \in X$, let $ntop(P, x)$ be the number of votes in P ranking x first. $P \sim_{r_P} P'$ holds if and only if for every x, $ntop(P, x) = ntop(P', x)$.*

Proof: The ($\Leftarrow$) direction is obvious. For the ($\Rightarrow$) direction, suppose there is an $x \in X$ such that $ntop(P, x) \neq ntop(P', x)$. Without loss of generality, assume $ntop(P, x) > ntop(P', x)$. Now, we have $\sum_{x \in P} ntop(P, x) = \sum_{x \in P'} ntop(P', x) = m$, therefore there must be an y such that $ntop(P, y) < ntop(P', y)$. Note that we necessarily have $y \neq x$. Now, let Q be the following profile with $2m - ntop(P, x) - ntop(P, y) + 1$ voters:

$m - ntop(P,x) + 1$ voters have x on top (and whatever below), and $m - ntop(P,y)$ voters have y on top (and whatever below). We have $ntop(P \cup Q, x) = m + 1$, $ntop(P \cup Q, y) = m$, and for every $z \neq x, y$, $ntop(P \cup Q, z) \leq m$. Therefore, $r_P(P \cup Q) = x$. Now, we have $ntop(P' \cup Q, x) = ntop(P', x) - ntop(P, x) + m + 1 \leq m$, $ntop(P' \cup Q, y) = ntop(P', y) - ntop(P, y) + m \geq m + 1$, and for every $z \neq x, y$, $ntop(P' \cup Q, z) \leq m$. Therefore, $r_P(P \cup Q) = y$. This shows that $P \not\sim_{r_P} P'$. ■

This characterization together with Proposition 1 tells us that the compilation complexity of r_P is exactly $\lceil \log L(m,p) \rceil$, where $L(m,p)$ be the number of vectors of positive integers $\langle \alpha_1, \ldots, \alpha_p \rangle$ such that $\sum_{i=1}^{p} \alpha_p = m$. The number of such vectors is known, in fact it is equivalent to the number of ways to choose m elements from a set of size p when repetition is allowed, that is $\binom{p+m-1}{m}$ —see *e.g.* [1]. A more explicit expression can be obtained at the price of a very tight approximation, by using Stirling's formula for factorials. The following result is then obtained after a few algebraic rewritings.

Corollary 1 *The compilation complexity of* r_P *is* $\Theta\left(p \log(1 + \frac{m}{p}) + m \log(1 + \frac{p}{m})\right)$

It can observed that the previous result yields an *upper bound* in $O(m + p)$, which can be compared with the "naive" upper bound that may be derived from the fact that it is sufficient to record the plurality scores of each candidate, which needs $O(p \log m)$ bits.

Borda. We get this intuitive characterization of $\sim$ for the Borda rule, in a similar way as Proposition 1 for plurality. More generally, a similar result holds for any scoring rule.

Lemma 2 *For* $P \in \mathcal{P}_X^m$ *and* $x \in X$*, let* $score_B(x, P)$ *be the Borda score of* x *obtained from the partial profile* P*.* $P \sim_{r_B} P'$ *holds if and only if for every* x*,* $score_B(x, P) = score_B(x, P')$*.*

Let us denote by $B(m,p)$ the number of vectors of positive integers $\langle \alpha_1, \ldots, \alpha_p \rangle$ corresponding to Borda scores once m votes have been expressed. Observe that we necessarily have that $\sum_{i=1}^{p} \alpha_p = \frac{mp(p-1)}{2}$, since each voter distributes $\frac{p(p-1)}{2}$ points among the candidates. However, this alone does not suffice to characterize the set of realizable Borda scores (for instance, if a candidate gets a score of 0, then no other candidate can get less than m). An upper bound is easily obtained by observing that is possible to simply record the scores of $p - 1$ candidates, and that this score can be at most $m(p-1)$.

Proposition 4 *The compilation complexity of Borda is at most* $(p-1) \log m(p-1)$*.*

Now we try to exhibit a lower bound that will approach this upper bound. The general idea is to restrict our attention to a subset of vectors of Borda scores. For example, for those vectors where the candidate with the lowest score gets between 0 and m, the second between m and $2m$, and so on until the penultimate voter, the score of the last candidate can be chosen on purpose so as to make a realizable vector of Borda scores. (Observe that by taking these intervals, the scores of the $p - 1$ first candidates can really be chosen independently).

In what follows, we show how to construct profiles that result in the desired vectors of Borda scores, albeit for the sake of readability we shall confine ourselves to a slightly more restricted case than the one discussed above. Technically, the bound obtained is slightly less tight, but the proof is easier to follow. Let us call *basic score* the vector of Borda scores obtained when all voters cast their vote similarly $\langle 0, 1, \ldots, p-1 \rangle$. The following Lemma shows that two voters can produce a vector where any candidate can obtain one more vote than the basic score, while the last candidate obtains one vote less.

Lemma 3 *For any $i < p$, the vector of Borda scores $\langle \alpha_1, \alpha_2, \ldots, \alpha_i + 1, \ldots, \alpha_p - 1\rangle$ where $\forall j \leq p, \alpha_j = 2(j-1)$ can result from a two-voter profile.*

Proof: We denote by $\langle \alpha_1^v, \alpha_2^v, \ldots, \alpha_n^v\rangle$ the vector corresponding to the ballot of voter v. We initially assign to voter 1 and 2 the basic vectors $\langle 0, 1, \ldots, p-1\rangle$. Now we construct the modified vectors of the two voters as follows: take the scores α_i^1 and α_{i+1}^1 of voter 1 and swap them; then take the scores α_{i+1}^2 and α_{i+2}^2 of voter 2 and swap them; then move back to voter 1 and swap the scores α_{i+2}^1 and α_{i+3}^1, and so on until the last score of voter 1 or voter 2 is reached, in which case no more swap is possible. Observe now that $\forall j \in [i+1, p-1]$, $\alpha_j^1 + \alpha_j^2 = \alpha_j'^1 + \alpha_j'^2$ because the swaps of voter 1 and 2 compensate each other, so the scores of these candidates remain unaffected. On the other hand, the Borda scores of candidate i and p are modified as required (resp. $+1$ and -1). ∎

But the same principle can be applied with m voters: in short, it is possible to distribute up to $m/2$ points among the first $p-2$ candidates to improve over their basic score (with the last candidate compensating by seeing its score decreased by the same amount of points):

Proposition 5 *Let $\{\delta_1, \ldots, \delta_{p-1}\}$ be any set of non-negative integers such that $\sum_{i=1}^{p-1} \delta_i \leq \frac{m}{2}$. The vector of Borda scores $\langle \delta_1, m+\delta_2, 2m+\delta_3, \ldots, 2(p-1)+\delta_p\rangle$, where $\delta_p = -\sum_{i=1}^{p-1} \delta_i$, can result from a m-voter profile.*

Proof: Let $m' = m - \sum_{i=1}^{p-1} 2\delta_i$. In the following, we will consider sums of profiles and multiplications by constants. In particular, $a \times \langle x_1, x_2 \ldots\rangle$ will refer to the profile $\langle ax_1, ax_2, \ldots\rangle$. The above profile can be decomposed as follows as a sum of scores

$$\begin{aligned}
\vec{\alpha}_1 &= 2\delta_1 \times \langle 0, 1, 2 \ldots\rangle + \langle \delta_1, 0, 0, 0, \ldots - \delta_1\rangle \\
\vec{\alpha}_2 &= 2\delta_2 \times \langle 0, 1, 2 \ldots\rangle + \langle 0, \delta_2, 0, 0, \ldots - \delta_2\rangle \\
\vec{\alpha}_3 &= 2\delta_3 \times \langle 0, 1, 2 \ldots\rangle + \langle 0, 0, \delta_3, 0, 0, \ldots - \delta_3\rangle \\
&\ldots \\
\vec{\alpha}_p &= m' \times \langle 0, 1, 2 \ldots\rangle
\end{aligned}$$

The last score can be realized by simply summing m' scores $\langle 0, 1, 2, \ldots\rangle$. As according to Lemma 3 the scores α_i can be obtained by summing $2\delta_i$ scores, the result follows. ∎

Corollary 2 *The compilation complexity of the Borda rule is $\Theta(p \log mp)$.*

Proof: Let $\mathbf{1_C}$ be the indicator function valued 1 if condition C is true and 0 otherwise. In Proposition 5, we showed that the number of profiles in which candidates $0 \ldots p-2$ have increasing scores is at least $\left|\left\{\langle \alpha_0 \ldots \alpha_{p-2}\rangle \in \mathbb{N}^{p-1} \mid \sum_{i=0}^{p-2} \alpha_i \leq \frac{m}{2}\right\}\right|$. More generally, the question amounts to enumerating V_t^s, the set of vectors of s non-negative integers, whose sum is lower or equal to t. This value can be written as $\int_{\alpha_0..\alpha_{s-1}=0}^{\infty} \mathbf{1}_{\sum \lfloor \alpha_i \rfloor \leq t} d\alpha_0 \ldots d\alpha_{s-1}$. Clearly, this can be lower bounded by $\int_0^{\infty} \mathbf{1}_{\sum \alpha_i \leq t} d\alpha_0 \ldots d\alpha_{s-1}$. But this is equal to half of the volume of the hypercube of dimension s whose side has length t. (For example, with $s = 2$, this value becomes half the area of a square $\frac{t^2}{2}$). More generally, we then have $V_t^s \geq \frac{t^s}{2}$. In our case, this gives us $\frac{1}{2} \times \left(\frac{m}{2}\right)^{p-1}$. Note that this lower bounds the number of profiles with increasing scores. Thus, the total number of profiles is at least $(p-1)!m^{p-1}2^{-p}$. Using the fact that $\log n! \geq n \log n$, we get the lower bound $(p-1)(\log_2(p-1) + \log_2 m - 2)$. Together with the upper bound, the result holds. ∎

3.2 Rules based on the weighted majority graph

We now consider tournament-based rules. Let P be a profile. $N_P(x,y)$ denotes the number voters in P preferring x to y. The *majority graph* M_P is the directed graph whose set of vertices is X and containing an edge from x to y if and only if $N_P(x,y) > N_P(y,x)$. The *weighted majority graph* $\mathcal{M}_P$ is the same as M_P, where each edge from x to y is weighted by $N(x,y)$ (note that there is no edge in $\mathcal{M}_P$ between x and y if and only if $N_P(x,y) = N_P(y,x)$.) A voting rule r is *based on the majority graph* (abridged into "MG-rule") if for any profile P, $r(P)$ can be computed from M_P, and *based on the weighted majority graph* (abridged into "WMG-rule") if for any profile P, $r(P)$ can be computed from $\mathcal{M}_P$. Obviously, a MG-rule is *a fortiori* a WMG-rule. A candidate x is the *Condorcet winner* for a profile P if it dominates every other candidate in M_P. A voting rule r is *Condorcet-consistent* if it elects the Condorcet winner whenever there exists one.

Lemma 4 *Let r be a WMG-rule rule. If $\mathcal{M}_P = \mathcal{M}_{P'}$ then $P \sim_r P'$.*

Proof: For any Q, $\mathcal{M}_{P\cup Q}$ is fully determined from $\mathcal{M}_P$ and $\mathcal{M}_Q$, because $N_{P\cup Q}(x,y) = N_P(x,y) + N_Q(x,y)$. If r is a WMG-rule then $r(P \cup Q)$ is fully determined from $\mathcal{M}_{P\cup Q}$, therefore from $\mathcal{M}_P$ and $\mathcal{M}_Q$, and a fortiori, from $\mathcal{M}_P$ and Q. ■

Note that for rules based on the (non-weighted) majority graph, we still need the *weighted* majority graph of P and P' to coincide – having only the majority graph coinciding is not sufficient for $P \sim_r P'$, since $M_{P\cup Q}$ is generally not fully determined from M_P and M_Q.

Lemma 4 gives an upper bound on the compilation complexity of a WMG-rule. Let $T(m,p)$ be the set of all weighted tournaments on X that can be obtained as the weighted majority graph of some m-voter profile.

Proposition 6 *If r is a WMG-rule then $C(r) \leq \log T(m,p)$.*

Getting a lower bound is not possible without a further assumption on r. After all, constant rules are based on the majority graph, yet they have a compilation complexity of 0. We say that a WMG-rule r is *proper* if $P \sim_r P'$ implies $\mathcal{M}_P = \mathcal{M}_{P'}$[2]. It is easy to find a natural sufficient condition for a WMG-rule to be proper:

Lemma 5 *If r is a Condorcet-consistent rule then $P \sim_r P'$ implies $\mathcal{M}_P = \mathcal{M}_{P'}$.*

Proof: Let r be a Condorcet-consistent rule. Assume $\mathcal{M}_P \neq \mathcal{M}_{P'}$, *i.e.*, there exists $(x,y) \in X$ with $N_P(x,y) \neq N_{P'}(x,y)$. W.l.o.g., $N_P(x,y) = N_{P'}(x,y) + k$ (hence $N_P(y,x) = N_{P'}(y,x) - k$), with $k > 0$. Let Q be a set of $m+1$ voters where: $m + 1 - N_P(x,y)$ voters prefer x to y and y to anyone else; $N_P(x,y)$ voters prefer y to x and x to anyone else. As we have $N_{P\cup Q}(x,y) = N_P(x,y) + N_Q(x,y) = m+1$; for any $z \neq x,y$, $N_{P\cup Q}(x,z) = N_P(x,z) + m + 1 \geq m+1$, x is Condorcet winner in $P \cup Q$ (which contains $2m+1$ voters) and $r(P \cup Q) = x$. But $N_{P'\cup Q}(y,x) = N_{P'}(y,x) + N_Q(y,x) = N_P(y,x) + k + N_P(x,y) = m + k$, and for any $z \neq x,y$, $N_{P'\cup Q}(y,z) = N_{P'}(y,z) + m + 1 \geq m+1$, so y is Condorcet winner in $P' \cup Q$ and $r(P' \cup Q) = y$. Hence $P \not\sim_r P'$. ■

This gives us the following lower bound.

[2]Examples of WMG-rules that are not proper: constant rules; dictatorial rules; strange rules such as $r(P)$ = first x_i (wrt a fixed ordering $x_1 > ... > x_p$ on candidates) such that for all $x_j \neq x_i$ there is at least one voter who prefers x_i to x_j, and x_p if there is no such x_i; "restricted" rules such as $r(P)$ being defined as the candidate maximizing the Copeland score among a fixed subset of candidates; etc.

Proposition 7 *If r is a Condorcet-consistent rule then $C(r) \geq \log T(m,p)$.*

From Propositions 6 and 7 we get

Proposition 8 *If r is a Condorcet-consistent WMG-rule, then $C(r) = \log T(m,p)$.*

Corollary 3 *The compilation complexity of the following rules is exactly $\log T(m,p)$:* Copeland, Simpson (maximin), Slater, Banks, uncovered set, Schwartz.

We now have to compute $T(m,p)$. We easily get the following upper bound.

Proposition 9 $\log T(m,p) \leq \frac{p(p-1)}{2} \log(m+2)$.

Proof: From Lemma 4 we know that it is enough to store M_P. Let $>$ be a fixed ordering on the candidates. Storing M_P can be done by storing, for every pair (x,y) of distinct candidates such that $x > y$, (a) a single bit indicating whether $N_P(x,y) > N_P(y,x)$ or $N_P(x,y) \leq N_P(y,x)$ and (b) $\min(N_P(x,y), N_P(y,x))$. Since the latter number can vary between 0 and $\frac{m}{2}$ if m is even, and between 0 and $\frac{m-1}{2}$ is m is odd, storing this number requires at most $\log\left(\frac{m}{2}+1\right)$ bits. This makes a total of $1 + \log\left(\frac{m}{2}+1\right)$ bits, that is, $\log(m+2)$ bits. We have $\frac{p(p-1)}{2}$ pairs of distinct candidates, hence the result. ∎

This bound is not necessarily reached: for any $x, y, z \in X$ and any profile P we have $N_P(x,z) \geq N_P(x,y) + N_P(y,z) - m$ (*e.g*, if $m = 3$ and $N_P(x,y) = N_P(y,z) = 2$, then $N_P(x,z)$ cannot be 0).

Lemma 6 *Consider V_t^s the set of vectors of s non-negative integers whose sum is lower or equal to t. Then $T(m,p) \geq \mid V_{\frac{m}{2}}^{\frac{p(p-1)}{2}} \mid$.*

Proof: Assume m is even. Let $\{c_{i,j} \mid 1 \leq i < j \leq p\}$ be any set non-negative integers such that $\sum_{i<j} c_{i,j} \leq \frac{m}{2}$. We will show how to build a profile such that $N(i,j) = 2c_{i,j}$, where $N(i,j)$ indicates how many voters prefer i to j. Let us divide voters into $\frac{p(p-1)}{2}$ groups $g_{i,j}$ with $1 \leq i < j \leq p$ and a final group g_0, such that each group $g_{i,j}$ is assumed to contain exactly $2c_{i,j}$ voters and g_0 contains the rest of the voters (i.e. $m - \sum_{i<j} 2c_{i,j}$). In each group $g_{i,j}$, set the profile of half of the voters to $i \succ j \succ x_1 \succ x_2 \succ \ldots \succ x_{p-2}$, and the other half to $x_{p-2} \succ x_{p-1} \succ \ldots \succ x_1 \succ i \succ j$, where $x_1 \ldots x_{p-2}$ refer to the candidates other than i and j in an arbitrary order. In group g_0 set half of the voters to $x_1 \succ x_2 \succ \ldots \succ x_p$ and the other half to $x_p \succ \ldots \succ x_1$. Let $N^g(x,y)$ denote the number of voters in group g preferring x to y. Clearly, $N^{g_{ij}}(x,y) = N(x,y)$ if $x = i$ and $y = j$; and 0 otherwise; and $N^{g_0}(x,y) = 0$. Thus, $N(x,y) = \sum N^{g_{i,j}}(x,y) = 2c_{i,j}$. ∎

From the previous Lemma, and using a technique similar to the one used in Corollary 2 to enumerate V_t^s, we obtain the compilation complexity of this family of rules:

Corollary 4 *If r is a Condorcet-consistent WMG-rule then $C(r) = \Theta(p^2 \log m)$.*

3.3 Plurality with runoff

Plurality with runoff is the voting rule (denoted by r_2) consisting of two rounds: the first round keeps only the two candidates with maximum plurality scores (with some tie-breaking mechanism), and the second round is simply the majority rule.

Proposition 10 *Let r_2 be the plurality-with-runoff rule. $P \sim_{r_2} Q$ holds if and only if for every x, $ntop(P,x) = ntop(Q,x)$ and for every x, y, $N_P(x,y) = N_Q(x,y)$.*

Lemma 7 *If for every x, $ntop(P,x) = ntop(Q,x)$ and for every x, y, $N_P(x,y) = N_Q(x,y)$, then $P \sim_{r_2} Q$.*

Proof: For every $x \in X$, since $ntop(P,x) = ntop(Q,x)$, we also have $ntop(P \cup R, x) = ntop(Q \cup R, x)$: the two plurality winners are the same in $P \cup R$ and $Q \cup R$. Let x and y be these two plurality winners. Since $N_P(x,y) = N_Q(x,y)$, we have $N_{P \cup R}(x,y) = N_{Q \cup R}(x,y)$, therefore, $M_{P \cup R}(x,y)$ if and only if $M_{Q \cup R}(x,y)$ and hence $r_2(P \cup R) = r_2(Q \cup R)$. ■

Lemma 8 *If for some x $ntop(P,x) \neq ntop(Q,x)$, then $P \not\sim_{r_2} Q$.*

Proof: If $p = 2$, this is a corollary of Lemma 1. Assume $p \geq 3$, and w.l.o.g., assume $ntop(P,x) > ntop(Q,x)$. Because $\sum_{c \in X} ntop(P,c) = \sum_{c \in X} ntop(Q,c) (= m)$, there exists an $y \neq x$ such that $ntop(P,y) < ntop(Q,y)$. Almost w.l.o.g., assume x has priority over y for tie-breaking[3]. Let $z \neq x, y$ (which is possible because $p \geq 3$). We now construct an R such that in $P \cup R$, the two finalists are x and z, and the winner is x, and in $Q \cup R$, the two finalists are y and z (therefore the winner cannot be x). Let R be the following partial profile containing $14m + ntop(P,y) - ntop(P,x)$ new votes:

$ntop(P,y) - ntop(P,x) + 4m$ votes:	$x \succ \ldots$
$4m$ votes:	$y \succ x \succ z \succ \ldots$
$6m$ votes:	$z \succ \ldots$

The plurality scores in $P \cup R$ are:

- $s_{P \cup R}(x) = ntop(P,x) + ntop(P,y) - ntop(P,x) + 4m = ntop(P,y) + 4m$;
- $s_{P \cup R}(y) = ntop(P,y) + 4m$;
- $s_{P \cup R}(z) = ntop(P,z) + 6m$.
- for every other candidate c, $s_{P \cup R}(c) = ntop(P,c)$.

Since $ntop(P,c) \leq m$ holds for every $c \neq x, y, z$, we have $s_{P \cup R}(z) > s_{P \cup R}(x) = s_{P \cup R}(y) > s_{P \cup R}(c)$ for every $c \neq x, y, z$. Because x has priority over y, the two candidates remaining for the second round are z and x. Now, the number of voters in $P \cup R$ preferring x to z is $N(P \cup R, x, z) = N(P,x,z) + ntop(P,y) - ntop(P,x) + 8m \geq 8m$ (because $N(P,x,z) \geq ntop(P,x)$); and $N(P \cup R, z, x) = N(P,z,x) + 6m \leq 7m$. Hence $r_2(P \cup R) = x$. The plurality scores in $Q \cup R$ are:

- $s_{Q \cup R}(x) = ntop(Q,x) + ntop(P,y) - ntop(P,x) + 4m > ntop(P,y) + 4m$;
- $s_{Q \cup R}(y) = ntop(Q,y) + 4m$;
- $s_{Q \cup R}(z) = ntop(Q,z) + 6m$.
- for every other candidate c, $s_{Q \cup R}(c) = ntop(Q,c)$.

[3]The proof in the opposite case is very similar and we omit it.

$s_{Q\cup R}(y) - s_{Q\cup R}(x) = ntop(Q,y) - ntop(P,y) + ntop(P,x) - ntop(Q,x)$. Now, by assumption we have $ntop(Q,y) > ntop(P,y)$ and $ntop(P,x) > ntop(Q,x)$, therefore $s_{Q\cup R}(y) > s_{Q\cup R}(x)$.

$s_{Q\cup R}(z) - s_{Q\cup R}(x) = ntop(Q,z) - ntop(Q,x) - ntop(P,y) + ntop(P,x) + 2m$. Now, $ntop(P,x) > ntop(Q,x)$, therefore $s_{Q\cup R}(z) - s_{Q\cup R}(x) > ntop(Q,z) - ntop(P,y) + 2m > 0$, that is, $s_{Q\cup R}(y) > s_{Q\cup R}(x)$.

Because the plurality scores of both y and z in $Q \cup R$ are larger than the plurality score of x, x does not pass the first round, therefore $r_2(P \cup R) \neq x$. ■

Lemma 9 *If for some $x, y \in X$, $N(P,x,y) \neq N(Q,x,y)$ then $P \not\sim_{r_2} Q$.*

Proof: Assume w.l.o.g. that $N(P,x,y) > N(Q,x,y)$. We are going to complete P and Q such that in both $P \cup R$ and $Q \cup R$, the finalists are x and y, with x winning in $P \cup R$ and y in $Q \cup R$. Let R be composed of the following $2N(P,y,x) + 3m + 1$ votes:

$2N(P,y,x) + m + 1$ votes:	$x \succ y \succ \ldots$
$2m$ votes:	$y \succ x \succ \ldots$

Obviously, the plurality scores in $P \cup R$ verify $s_{P\cup R}(x) > m$, $s_{P\cup R}(y) > m$, and for any $c \neq x, y$, $s_{P\cup R}(c) \leq m$, therefore, the finalists are x and y. Things are the same for $Q \cup R$.

Now, $N(P \cup R, x, y) = N(P,x,y) + 2N(P,y,x) + m + 1 = m + N(P,y,x) + m + 1 = N(P,y,x) + 2m + 1$; and $N(P \cup R, x, y) = 2N(P,y,x) + 4m + 1 - N(P \cup R, x, y) = N(P,y,x) + 2m$. Therefore, $r_2(P \cup R) = x$.

Lastly, $N(Q \cup R, x, y) = N(Q,x,y) + 2N(P,y,x) + m + 1$, and $N(Q \cup R, x, y) = N(Q,y,x) + 2m$. We have now $N(Q \cup R, y, x) - N(Q \cup R, x, y) = N(Q,y,x) + m - N(Q,x,y) - 2N(P,y,x) - m - 1 = N(Q,y,x) + m - N(Q,x,y) - 2N(P,x,y) - 1 = 2(N(Q,y,x) - N(P,y,x)) - 1$. Now, $N(Q,y,x) > N(P,y,x)$, therefore, $N(Q \cup R, y, x) > N(Q \cup R, x, y)$, that is, $r_2(Q \cup R) = y$. ■

Proposition 10 is now a corollary from Lemmas 7, 8 and 9, and it follows that:

Proposition 11 *The compilation complexity of plurality with runoff is* $\log L(m,p) + \log T(m,p)$.

4 Conclusion

This paper has introduced a notion that we believe to be of primary importance in many practical situations: the compilation of incomplete profiles. In particular, the amount of information that a given polling station needs to transmit to the central authority is a good indicator of the difficulty of the verification process. We have established a general technique which allows us to derive the compilation complexity of a voting rule, and have related it to other issues in communication complexity. We have derived a number of results for specific classes of voting rules. A question that we have only sketched in this paper and that we plan to consider more carefully concerns the situations where the number k of remaining voters is fixed. In this case, a different approach can be taken: instead of compiling the partial profiles as provided by the m voters, it may be more efficient to compile the possible completion of this partial profile together with the associated outcome, or, in other words, to compile the function that takes the remaining k profiles as input (there are $(p!)^k$ such inputs) and return the outcome. As there are $(p!)^k$ possible profiles, the number of such functions is $p^{(p!)^k}$. This tells us that a general upper bound for $C(r,k)$ is $\leq (p!)^k . \log p$.

Hence, overall we have $C(r,k) \leq \min(m.\log(p!), (p!)^k.\log p)$ (note that the second term becomes interesting only when m is big enough, and p and k small enough).

The general problem of dealing with incomplete profiles opens a host of related questions, for instance the probability that a voting process could be stopped after only m voters have expressed their opinions, or (a closely related question), the probability that the central authority would make a mistake were it forced to commit on a winner in situations where no candidate is yet guaranteed to prevail. Another interesting issue for further research would consist in designing new ways of computing NP-hard voting rules using an off-line compilation step so that their on-line computation time becomes polynomial in the size of the initial profile. This, of course, implies that the size of $\sigma(P)$ may be much larger (possibly exponentially) than the size of the input, which means that the computation of ρ may be logarithmic in the size of the compilation $\sigma(P)$.

References

[1] E. Bender and S. Williamson. *The Foundations of Combinatorics with Applications*. Dover, 2006.

[2] S. Brams, D. Kilgour, and W. Zwicker. The paradox of multiple elections. *Social Choice and Welfare*, 15(2):211–236, 1998.

[3] M. Cadoli, F. M. Donini, P. Liberatore, and M. Schaerf. Feasibility and unfeasibility of off-line processing. In *ISTCS*, 1996.

[4] V. Conitzer and T. Sandholm. Vote elicitation: complexity and strategy-proofness. In *Proceedings of AAAI-02*, pages 392–397, 2002.

[5] V. Conitzer and T. Sandholm. Communication complexity of common voting rules. In *Proceedings of EC-05*, 2005.

[6] A. Darwiche and P. Marquis. A knowledge compilation map. *JAIR*, 17:229–264, 2002.

[7] K. Konczak and J. Lang. Voting procedures with incomplete preferences. In *Proc. IJCAI-05 Multidisciplinary Workshop on Advances in Preference Handling*, 2005.

[8] E. Kushilevitz and N. Nisan. *Communication complexity*. Cambridge Univ. Press, 1997.

[9] M.S. Pini, F. Rossi, K. Brent Venable, and T. Walsh. Incompleteness and incomparability in preference aggregation. In *IJCAI*, 2007.

[10] T. Walsh. Complexity issues in preference elicitation and manipulation. In *AAMAS*, 2008.

[11] L. Xia and V. Conitzer. Determining possible and necessary winners under common voting rules given partial orders. In *Proceedings of AAAI-08*, 2008.

Y. Chevaleyre, N. Maudet, G. Ravilly-Abadie
LAMSADE, Univ. Paris-Dauphine
75775 Paris, France
Email: {chevaleyre,maudet,ravilly-abadie}@lamsade.dauphine.fr

J. Lang
IRIT, CNRS
31062 Toulouse, France
Email: lang@irit.fr

Preference Functions That Score Rankings and Maximum Likelihood Estimation

Vincent Conitzer Matthew Rognlie Lirong Xia

Abstract

A preference function (PF) takes a set of votes (linear orders over a set of alternatives) as input, and produces one or more rankings (also linear orders over the alternatives) as output. Such functions have many applications, for example, aggregating the preferences of multiple agents, or merging rankings (of, say, webpages) into a single ranking. The key issue is choosing a PF to use. One natural and previously studied approach is to assume that there is an unobserved "correct" ranking, and the votes are noisy estimates of this. Then, we can use the PF that always chooses the maximum likelihood estimate (MLE) of the correct ranking. In this paper, we define simple ranking scoring functions (SRSFs) and show that the class of neutral SRSFs is exactly the class of neutral PFs that are MLEs for some noise model. We also define extended ranking scoring functions (ERSFs) and show a condition under which these coincide with SRSFs. We study key properties such as consistency and continuity, and consider some example PFs. In particular, we study Single Transferable Vote (STV), a commonly used PF, showing that it is an ERSF but not an SRSF, thereby clarifying the extent to which it is an MLE function. This also gives a new perspective on how ties should be broken under STV. We leave some open questions.

1 Introduction

In a typical social choice setting, there is some set of alternatives, and multiple rankings of these alternatives are provided. These input rankings are called the *votes*. Based on these votes, the goal is either to choose one alternative, or to create an aggregate ranking of all the alternatives. In this paper, we will be interested in the latter goal; if it is desired to choose one alternative, then we can simply choose the top-ranked alternative in the aggregate ranking. Formally, a *preference function (PF)*[1] takes a set of votes (linear orders over the alternatives) as input, and produces one or more aggregate rankings (also linear orders over the alternatives) as output. (The only reason for allowing multiple aggregate rankings is to account for the possibility of ties.)

The key issue is to choose a rule for determining the aggregate ranking, that is, a preference function. So, we may ask the following (vague) question: *What is the optimal preference function?* This has been (and will likely continue to be) a topic of debate for centuries among social choice theorists. Many different PFs have been proposed, each with its own desirable properties; some of them have elegant axiomatizations. Presumably, which PF is optimal depends on the setting at hand. For example, in some settings, the voters are agents that each have their own idiosyncratic preferences over the alternatives, and the only purpose of voting is to reach a compromise. In such a setting, no alternative can be said to be better than another alternative in any *absolute* sense: an alternative's quality is defined relative to the votes. In such a setting, it makes sense to pay close attention to issues such as the manipulability of the PF.

In other settings, however, there is more of an absolute sense in which some alternatives are better than others. For example, when we wish to aggregate rankings of webpages, provided by multiple search engines in response to the same query, it is reasonable to believe that some of these pages are in fact more relevant than others. The reason that not all of the search engines agree on

[1]We use "preference function" rather than "social welfare function" because the resulting set of strict rankings need not correspond to a weak ranking (where a set of strict rankings "corresponds" to a weak ranking if it consists of all the strict rankings that can be obtained by breaking the ties in the weak ranking). The term "preference function" has previously been used in this context [14].

the ranking is that the search engines are unable to directly perceive this absolute relevance of the pages. Here, it makes sense to think of each vote as a *noisy estimate* of the correct, absolute ranking. Our goal is to find an aggregate ranking that is as close as possible to the correct ranking, based on these noisy estimates. This is the type of setting that we will study in this paper.

In a 2005 paper, Conitzer and Sandholm considered the following way of making this precise [3]. There is a correct ranking r of the alternatives; given r, for every ranking v, there is a conditional probability $P(v|r)$ that a given voter will cast vote v. (In this paper, we do not consider the possibility that different voters' votes are drawn according to different conditional distributions.) Votes are conditionally independent given r. Put another way, the noise that each voter experiences is i.i.d. The Bayesian network in Figure 1 illustrates this setup.

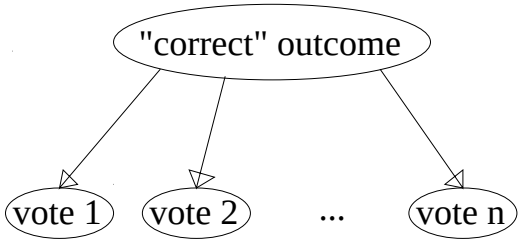


Figure 1: A Bayesian network representation.

The votes are the observed variables, and the noise that a voter experiences is represented by the conditional probability table of that vote. Under this setup, a natural goal is to find the maximum likelihood estimate (MLE) of the correct ranking. (If r is drawn uniformly at random, this maximum likelihood estimate also maximizes the posterior probability.) The function that takes the votes as input and produces the MLE ranking(s) as output is a preference function; in a sense, it is the optimal one for the particular noise model at hand.

As pointed out by Conitzer and Sandholm, they were not the first to consider this type of setup. In fact, the basic idea dates back over two centuries to Condorcet [4], who studied one particular noise model. He solved for the MLE PF for two and three alternatives under this model; the general solution was given two centuries later by Young [13], who showed that the MLE PF for Condorcet's model coincides with a function proposed by Kemeny [7]. This has frequently been used as an argument in favor of using Kemeny's PF; however, different noise models will in general result in different MLE PFs. Several generalizations of this basic noise model have been studied [6, 5, 8, 9]. Conitzer and Sandholm considered the opposite direction: they studied a number of specific well-known PFs and they showed that for some of them, there exists a noise model such that this PF becomes the MLE, whereas for others, no such noise model can be constructed. This shows that the former PFs are in a sense more natural than the latter. Also, when a noise model can be constructed, it gives insight into the PF; moreover, if the noise model is unreasonable in a certain way, it can be modified, resulting in an improved PF.

In this paper, we continue this line of work. We provide an exact characterization of the class of (neutral) PFs for which a noise model can be constructed: we show that this class is equal to the class of (neutral) *simple ranking scoring functions (SRSFs)*, which, for every vote, assign a score to every potential aggregate ranking, and the ranking(s) with the highest total score win(s). We show that several common PFs are SRSFs (these proofs resemble the corresponding proofs by Conitzer and Sandholm that these PFs are MLEs, but the proofs are significantly simpler in the language of SRSFs). We also consider *extended ranking scoring functions (ERSFs)*, which coincide with SRSFs except they can break ties according to another SRSF, and remaining ties according to another SRSF, *etc.* We show that if there is a bound on the number of votes, then the two classes (SRSFs and ERSFs) coincide. We study some basic properties of SRSFs and ERSFs, some of them closely related to Conitzer and Sandholm's proof techniques. Finally, we study one PF, Single Transferable

Vote (STV), also known as Instant Runoff Voting, in detail. STV is used in many elections around the world; additionally, it illustrates a number of key points about our results. A noise model for STV was given by Conitzer and Sandholm. However, this noise model involves probabilities that are infinitesimally smaller than other probabilities. We show that such infinitesimally small probabilities are in a sense necessary, by showing that STV is in fact not an SRSF (when there is no bound on the number of votes). Still, we do show that STV is an ERSF (in a way that resembles the noise model with infinitesimally small probabilities). Hence, STV is in fact an MLE PF if there is an upper bound on the number of votes. Along the way, some interesting questions arise about how ties should be broken under STV. We propose two ways of breaking ties that we believe are perhaps more sensible than the common way, although at least one of the ways leads to computational difficulties. We also leave some open questions.

2 Definitions

In the below, we let A be the set of alternatives, $|A| = m$, and $L(A)$ the set of linear orders over (that is, strict rankings of) these alternatives. A *preference function (PF)* is a function $f : \bigcup_{i=0,1,2,\ldots} L(A)^i \rightarrow 2^{L(A)} - \emptyset$. That is, f takes as input a vector (of any length) V of linear orders (votes) over the alternatives, and as output produces one or more linear orders over (aggregate rankings of) the alternatives. (On many inputs, only a single ranking is produced, but it is possible that there are ties.) Input vectors are also called *profiles*. We restrict our attention to PFs that are *anonymous*, that is, they treat all votes equally; hence, a profile can be thought of as a multiset of votes. Below are the PFs that we will study in this paper.

- *Positional scoring functions.* A positional scoring function is defined by a vector $(s_1, \ldots, s_m) \in \mathbb{R}^m$, with $s_1 \geq s_2 \geq \ldots \geq s_m$. An alternative receives s_i points every time it is ranked ith. Alternatives are ranked by how many points they receive; if some alternatives end up tied, then they can be ranked in any order (and all the complete rankings that can result from this will be produced by the PF). Examples include *plurality* ($s_1 = 1, s_2 = s_3 = \ldots = s_m = 0$), *veto* or *anti-plurality* ($s_1 = s_2 = \ldots = s_{m-1} = 1, s_m = 0$), and *Borda* ($s_1 = m-1, s_2 = m-2, \ldots, s_m = 0$).

- *Kemeny.* Given a vote v, a possible ranking r, and two alternatives a, b, let $\delta(v, r, a, b) = 1$ if $a \succ_v b$ and $a \succ_r b$, and $\delta(v, r, a, b) = 0$ otherwise. Then, $f(V) = \arg\max_{r \in L(A)} \sum_{a,b \in A} \sum_{v \in V} \delta(v, r, a, b)$. That is, we choose the ranking(s) that maximize(s) the total number of times that the ranking agrees with a vote on a pair of alternatives.

- *Single Transferable Vote (STV).* The alternative with the lowest plurality score (that is, the one that is ranked first by the fewest votes) is ranked last, and is removed from all the votes (so that the plurality scores change). The remainder of the ranking is determined recursively. (We will have more to say about how ties are broken later.)

A PF is *neutral* if treats all alternatives equally. To be precise, a PF is neutral if for any votes V and any permutation π on the alternatives, $f(\pi(V)) = \pi(f(V))$. Here, a permutation is applied to a vector or set of rankings of the alternatives by applying it to each individual alternative in those rankings. Naturally, neutrality is a common requirement. Another common requirement for an anonymous PF is *homogeneity*: if we multiply the profile by some natural number $n > 0$ (that is, replace each vote by n duplicates of it), then the outcome should not change. All of the above PFs are anonymous, neutral, and homogenous.

We now define noise models and MLE PFs formally.

Definition 1 *A* noise model ν *specifies a probability* $P_\nu(v|r)$ *for every* $v, r \in L(A)$.

Definition 2 *A noise model ν is* neutral *if for any v, r, and permutation π on A, we have $P_\nu(v|r) = P_\nu(\pi(v)|\pi(r))$.*

Definition 3 *A PF f is a* maximum likelihood estimator (MLE) *if there exists a noise model ν so that $f(V) = \arg\max_{r \in L(A)} \prod_{v \in V} P_\nu(v|r)$.*

We now define simple ranking scoring functions. Effectively, every vote gives a number of points to every possible aggregate ranking, and the ranking(s) with the most points win(s).

Definition 4 *A PF f is a* simple ranking scoring function (SRSF) *if there exists a function $s : L(A) \times L(A) \to \mathbb{R}$ such that for all V, $f(V) = \arg\max_{r \in L(A)} \sum_{v \in V} s(v, r)$.*

Definition 5 *A function $s : L(A) \times L(A) \to \mathbb{R}$ is* neutral *if for any v, r, and permutation π on A, $s(v, r) = s(\pi(v), \pi(r))$.*

An SRSF can be run by explicitly computing each ranking's score, but because there are $m!$ rankings this is impractical for all but the smallest numbers of alternatives. However, such explicit computation is generally not necessary. For example, we will see that positional scoring functions as well as the Kemeny function are SRSFs. Positional scoring functions are of course easy to run; running the Kemeny function is in fact NP-hard [1], but can in practice be done quite fast [2, 9].

3 Equivalence of neutral MLEs and SRSFs

We now show the equivalence of MLEs and SRSFs. We only show this for neutral PFs; in fact, it is not true for PFs that are not neutral. For example, a PF that always chooses the same ranking r^* regardless of the votes is an SRSF, simply by setting $s(v, r^*) = 1$ for all v and setting $s(v, r) = 0$ everywhere else. However, this PF is not an MLE: given a noise model ν, if we take another ranking $r \neq r^*$, we must have $\sum_{v \in L(A)} P_\nu(v|r) = 1 = \sum_{v \in L(A)} P_\nu(v|r^*)$, hence there exists some v such that $P_\nu(v|r) \geq P_\nu(v|r^*)$; it follows that r^* is not the (sole) winner if v is the only vote.

Lemma 1 *A neutral PF f is an MLE if and only if it is an MLE for a neutral noise model.*

Proof: The "if" direction is immediate. For the "only if" direction, given a noise model ν for f, construct a new noise model ν' as follows: $P_{\nu'}(v|r) = (1/m!) \sum_\pi P_\nu(\pi(v)|\pi(r))$. (Here, π ranges over permutations of A.) This is still a valid noise model because $\sum_{v \in L(A)} P_{\nu'}(v|r) = \sum_{v \in L(A)} (1/m!) \sum_\pi P_\nu(\pi(v)|\pi(r)) = (1/m!) \sum_\pi \sum_{v \in L(A)} P_\nu(\pi(v)|\pi(r)) = 1$. ν' is also neutral because $P_{\nu'}(\pi(v)|\pi(r)) = (1/m!) \sum_{\pi'} P_\nu(\pi'(\pi(v))|\pi'(\pi(r))) = (1/m!) \sum_{\pi''} P_\nu(\pi''(v)|\pi''(r)) = P_{\nu'}(v|r)$. Also, if $r^* \in \arg\max_{r \in L(A)} \prod_{v \in V} P_\nu(v|r)$, then by the neutrality of f, for any π, $\pi(r^*) \in \arg\max_{r \in L(A)} \prod_{v \in V} P_\nu(\pi(v)|r)$. Hence, $r^* \in \arg\max_{r \in L(A)} (1/m!) \sum_\pi \prod_{v \in V} P_\nu(\pi(v)|\pi(r)) = \arg\max_{r \in L(A)} \prod_{v \in V} (1/m!) \sum_\pi P_\nu(\pi(v)|\pi(r)) = \arg\max_{r \in L(A)} \prod_{v \in V} P_{\nu'}(v|r)$. Conversely, it can similarly be shown that if $r^* \notin \arg\max_{r \in L(A)} \prod_{v \in V} P_\nu(v|r)$, then $r^* \notin \arg\max_{r \in L(A)} \prod_{v \in V} P_{\nu'}(v|r)$. Hence, ν' is a valid noise model for f. ∎

Lemma 2 *A neutral PF f is an SRSF if and only if it is an SRSF for a neutral function s'.*

Proof: The "if" direction is immediate. For the "only if" direction, given a function s, construct a new function s' as follows: $s'(v, r) = \sum_\pi s(\pi(v), \pi(r))$. s' is neutral because $s'(\pi(v), \pi(r)) = \sum_{\pi'} s(\pi'(\pi(v)), \pi'(\pi(r))) = \sum_{\pi''} s(\pi''(v), \pi''(r)) = s'(v, r)$. Also, if $r^* \in \arg\max_{r \in L(A)} \sum_{v \in V} s(v, r)$, then by the neutrality of f, for any π, $\pi(r^*) \in \arg\max_{r \in L(A)} \sum_{v \in V} s(\pi(v), r)$. Hence, $r^* \in \arg\max_{r \in L(A)} \sum_\pi \sum_{v \in V} s(\pi(v), \pi(r)) =$

$\arg\max_{r \in L(A)} \sum_{v \in V} \sum_{\pi} s(\pi(v), \pi(r)) = \arg\max_{r \in L(A)} \sum_{v \in V} s'(v, r)$. Conversely, it can similarly be shown that if $r^* \notin \arg\max_{r \in L(A)} \sum_{v \in V} s(v, r)$, then $r^* \notin \arg\max_{r \in L(A)} \sum_{v \in V} s'(v, r)$. Hence, s' is a valid function for f. ■

We can now prove the characterization result:

Theorem 1 *A neutral PF is an MLE if and only if it is an SRSF.*

Proof: If f is an MLE, then for some neutral ν, $f(V) = \arg\max_{r \in L(A)} \prod_{v \in V} P_\nu(v|r) = \arg\max_{r \in L(A)} \log(\prod_{v \in V} P_\nu(v|r)) = \arg\max_{r \in L(A)} \sum_{v \in V} \log(P_\nu(v|r))$. Hence it is the SRSF where $s(v, r) = \log(P_\nu(v|r))$ (here, s is neutral).

Conversely, if f is an SRSF, then for some neutral s, $f(V) = \arg\max_{r \in L(A)} \sum_{v \in V} s(v, r) = \arg\max_{r \in L(A)} 2^{\sum_{v \in V} s(v,r)} = \arg\max_{r \in L(A)} \prod_{v \in V} 2^{s(v,r)}$. Because s is neutral, we have that $\sum_{v \in L(A)} 2^{s(v,r)}$ is the same for all r. (This is because for any r_1, r_2, there exists a permutation π on A such that $\pi(r_1) = r_2$, so that we have $\sum_{v \in L(A)} 2^{s(v,r_1)} = \sum_{v \in L(A)} 2^{s(\pi(v),r_2)}$ by neutrality, which by changing the order of the summands is equal to $\sum_{v \in L(A)} 2^{s(v,r_2)}$.) It follows that $f(V) = \arg\max_{r \in L(A)} \prod_{v \in V} (2^{s(v,r)})/(\sum_{v' \in L(A)} 2^{s(v',r)})$. Hence f is the maximum likelihood estimator for the noise model ν defined by $P_\nu(v|r) = (2^{s(v,r)})/(\sum_{v' \in L(A)} 2^{s(v',r)})$. ■

4 Examples of SRSFs

We now show that some common PFs are SRSFs. These proofs resemble the corresponding proofs by Conitzer and Sandholm that these functions are MLEs, but they are simpler. These propositions also follow from the work of Zwicker [15].

Proposition 1 *Every positional scoring function is an SRSF.*

Proof: Given a positional scoring function, let $t : L(A) \times A \to \mathbb{R}$ be defined as follows: $t(v, a)$ is the number of points that a gets for vote v. Then, let $s(v, r) = \sum_{i=1}^{m} (m - i) t(v, r(i))$, where $r(i)$ is the alternative ranked ith in r. Let us consider the SRSF defined by this function s; it selects $\arg\max_{r \in L(A)} \sum_{v \in V} s(v, r) = \arg\max_{r \in L(A)} \sum_{v \in V} \sum_{i=1}^{m} (m - i) t(v, r(i)) = \arg\max_{r \in L(A)} \sum_{i=1}^{m} (m - i) \sum_{v \in V} t(v, r(i))$. Here, $\sum_{v \in V} t(v, r(i))$ is the total score that alternative $r(i)$ receives under the positional scoring function. Because $m - i$ is decreasing in i, to maximize $\sum_{i=1}^{m} (m - i) \sum_{v \in V} t(v, r(i))$, we should rank the alternative with the highest total score first, the one with the next-highest total score second, *etc.* If some of the alternatives are tied, they can be ranked in any order. ■

Not only positional scoring functions are SRSFs, however.

Proposition 2 *The Kemeny PF is an SRSF.*

Proof: This is almost immediate: we defined the Kemeny PF by $f(V) = \arg\max_{r \in L(A)} \sum_{a,b \in A} \sum_{v \in V} \delta(v, r, a, b)$, so we simply let $s(v, r) = \sum_{a,b \in A} \delta(v, r, a, b)$. ■

5 Extended ranking scoring functions

An *extended ranking scoring function (ERSF)* starts by running an SRSF, then (potentially) breaks ties according to another SRSF, and (potentially) any remaining ties according to yet another SRSF, *etc.* Formally:

Definition 6 *An* ERSF f of depth k *consists of an ERSF* f' *of depth* $k-1$ *and a function* $s_d : L(A) \times L(A) \to \mathbb{R}$. *It chooses* $f(V) = \arg\max_{r \in f'(V)} \sum_{v \in V} s_d(v, r)$. *An ERSF of depth* 0 *returns the set of all rankings* $L(A)$.

So, an ERSF of (finite) depth d is defined by a sequence $f_1, \ldots, f_d$ of SRSFs. We can think of the scores at each depth as being infinitesimally smaller than the ones at the previous depths. We can multiply the scores at depth l by ϵ^l for some small ϵ and then add all the scores together to obtain an SRSF; however, this SRSF will in general be different from the ERSF. Nevertheless, if ϵ is small relative to the number of votes, then the two will coincide. This is the intuition behind the following result:

Proposition 3 *For any ERSF, for any natural number* N*, there exists an SRSF that agrees with the ERSF as long as there are at most* N *votes.*

Proof: Let the sequence of SRSFs $f_1, \ldots, f_d$, defined by scoring functions $s_1, \ldots, s_d$, define the ERSF f; we prove the claim by induction. The claim is trivial for $d = 1$. Let us assume that we have proven the result for $d = k - 1$; we will show it for $d = k$. Let f' be the ERSF corresponding to the first $d - 1$ SRSFs, and, by the induction assumption, let s define the SRSF that agrees with f' when there are at most N votes. There are only finitely many profiles V of size at most N; hence, there must be some ϵ such that $\sum_{v \in V} s(v, r) < \sum_{v \in V} s(v, r')$ and $|V| \leq N$ implies that $\sum_{v \in V} s(v, r) + \epsilon < \sum_{v \in V} s(v, r')$. Now let us consider s_d; there must exist some $H \in \mathbb{R}$ such that $|V| \leq N$ implies $\sum_{v \in V} s_d(v, r) < H$. Then, let s' be defined by $s'(v, r) = s(v, r) + (\epsilon / H) s_d(v, r)$. On profiles of size at most N, the second term will contribute at most ϵ to the total score of any r, so if r receives a strictly lower total score than r' under s, it will also receive a strictly lower score under s'. Hence, the only effect of the second term is to break ties according to s_d; so the SRSF defined by s' coincides with the original ERSF f when there are at most N votes. ∎

Thus, for all practical purposes, we can simulate an ERSF with an SRSF. (Of course, every SRSF is also an ERSF.)

6 Properties of SRSFs and ERSFs

In this section, we study some important properties of SRSFs and ERSFs. Specifically, we study *consistency* and *continuity*. There are several related works that study similar properties and derive related results, but there are significant differences in the setup. Smith [11] and Young [12] study these properties in *social choice rules*, which select one or more alternatives as the winner(s); we will discuss their results in more detail in Section 8. However, consistency in the context of preference functions (studied previously by Young and Levenglick [14]) is significantly different from consistency in the context of social choice rules. Other related work includes Myerson [10], who extends the Smith and Young result to settings where voters do not necessarily submit a ranking of the alternatives, and Zwicker [15], who studies a general notion of scoring rules and shows these rules are equivalent to *mean proximity rules*, which compute the mean location of the votes according to some embedding in space, and then choose the closest outcome(s).

An anonymous PF f is *consistent* if for any pair of profiles V_1 and V_2, if $f(V_1) \cap f(V_2) \neq \emptyset$, then $f(V_1 + V_2) = f(V_1) \cap f(V_2)$ (where addition is defined in the natural way). That is, if the rankings that f produces given V_1 overlap with those that f produces given V_2, then when V_1 and V_2 are taken together, f must produce the rankings that were produced in both cases, and no others.

Proposition 4 *Any ERSF is consistent.*

Proof: Let f be an ERSF of depth k, defined by a sequence of SRSFs $f_1, \ldots, f_k$ with score functions $s_1, \ldots, s_k$. For any $i \leq k$, let F_i be the ERSF of depth i defined by the sequence $f_1, \ldots, f_i$. Let V_1,

V_2 be profiles such that $f(V_1) \cap f(V_2) \neq \emptyset$; this also implies that $F_i(V_1) \cap F_i(V_2) \neq \emptyset$ for all $i \leq k$. We use induction on i to prove that for any $i \leq k$, $F_i(V_1 + V_2) = F_i(V_1) \cap F_i(V_2)$. When $i = 1$, $F_1(V_1) = f_1(V_1)$ is the set of rankings r that maximize $s_1(V_1, l)$; $F_1(V_2) = f_1(V_2)$ is the set of rankings r that maximize $s_1(V_2, l)$. Therefore, $F_1(V_1) \cap F_1(V_2)$ (which we know is nonempty) is the set of rankings r that maximize $s_1(V_1 + V_2, r)$. Now, suppose that for some $i \leq k$, $F_i(V_1 + V_2) = F_i(V_1) \cap F_i(V_2)$. $F_{i+1}(V_1)$ ($F_{i+1}(V_2)$) is the set of rankings $r \in F_i(V_1)$ ($r \in F_i(V_2)$) that maximize $s_{i+1}(V_1, r)$ ($s_{i+1}(V_2, r)$). Hence, $F_{i+1}(V_1) \cap F_{i+1}(V_2)$ (which we know is nonempty) is the set of rankings $r \in F_i(V_1) \cap F_i(V_2)$ that maximize $s_{i+1}(V_1, r) + s_{i+1}(V_2, r) = s_{i+1}(V_1 + V_2, r)$. By the induction assumption, we have that $F_i(V_1) \cap F_i(V_2) = F_i(V_1 + V_2)$, and we know that the set of rankings $r \in F_i(V_1 + V_2)$ that maximize $s_{i+1}(V_1 + V_2, r)$ is equal to $F_{i+1}(V_1 + V_2)$. It follows that $F_{i+1}(V_1) \cap F_{i+1}(V_2) = F_{i+1}(V_1 + V_2)$, completing the induction step. When $i = k$, $F_k = f$, which completes the proof. ∎

The proofs by Conitzer and Sandholm [3] that several PFs are not MLEs effectively come down to showing examples where these PFs are not consistent. By the above result, this implies that they are not ERSFs (and hence not SRSFs, and hence not MLEs). Formally (we will not define these PFs in this paper):

Proposition 5 *The Bucklin, Copeland, maximin, and ranked pairs PFs are not ERSFs.*

Proof: None of these PFs are consistent: counterexamples can be found in the proofs of Conitzer and Sandholm [3]. ∎

Let $L(A) = \{l_1, \ldots, l_{m!}\}$. For any anonymous PF f, any profile V can be rewritten as a linear combination of the linear orders in $L(A)$. Let $V = \sum_{i=1}^{m!} t_i l_i$, where for any $i \leq m!$, t_i is a non-negative integer. If f is also homogenous, then the domain of f can be extended to the set of all *fractional profiles* $V = \sum_{i=1}^{m!} t_i l_i$ where each t_i is a nonnegative rational number, as follows. We choose $N_V > 0, N_V \in \mathbb{N}$ such that for every $i \leq m!$, $t_i N_V$ is a integer. Then, we let $f(V) = f(N_V V)$. (This is well-defined because of the homogeneity.)

A fractional profile V can be viewed as a point in the $m!$-dimensional space $(\mathbb{Q}^{\geq 0})^{m!}$ where the coefficient t_i is the component of the ith dimension. Thus, in a slight abuse of notation, we can apply f to vectors of $m!$ nonnegative rational numbers, under the interpretation that $f(t_1, \ldots, t_{m!}) = f(\sum_{i=1}^{m!} t_i l_i)$. The extension of f to $(\mathbb{Q}^{\geq 0})^{m!}$ allows us to define continuity. An anonymous PF f is *continuous* if for any sequence of points $p_1, p_2, \ldots \in (\mathbb{Q}^{\geq 0})^{m!}$ with 1. $\lim_{i \to \infty} p_i = p$, and 2. for all $i \in \mathbb{N}$, $r \in f(p_i)$, we have $r \in f(p)$. That is, if f produces some ranking r on every point along a sequence that converges to a limit point, then f should also produce r at the limit point.[2]

Proposition 6 *Any SRSF is continuous.*

Proof: For any sequence of points $p_1, p_2, \ldots \in (\mathbb{Q}^{\geq 0})^{m!}$ with $\lim_{i \to \infty} p_i = p$, we have that for all $r \in L(A)$, $\lim_{i \to \infty} s(p_i, r) = s(p, r)$. If $r \in f(p_i)$ for all i, then for any $r' \in L(A)$, $s(p_i, r) \geq s(p_i, r')$, hence we have $s(p, r) = \lim_{i \to \infty} s(p_i, r) \geq \lim_{i \to \infty} s(p_i, r') = s(p, r')$. It follows that $r \in f(p)$. ∎

In contrast, ERSFs are not necessarily continuous, as shown by the following example. Let f_1 be the SRSF defined by the score function s_1, which is defined by $s_1(v, r) = 1$ if $v = r$ and $s_1(v, r) = 0$ if $v \neq r$. Let f_2 be the Borda function. Let f be the ERSF defined by the sequence f_1, f_2. Let $m = 3$ with alternatives A, B, and C, and let $p = \{A \succ B \succ C, B \succ C \succ A, C \succ B \succ A\}$. We have $f(p) = \{B \succ C \succ A\}$, but for any $\epsilon > 0$, $f(p + \epsilon(A \succ B \succ C)) = f_1(p + \epsilon(A \succ B \succ$

[2]Our definition of continuity is equivalent to the correspondence being *upper hemicontinuous*, or *closed* (the two are equivalent in this context).

$C)) = \{A \succ B \succ C\}$. Therefore, if we let $p_i = p + \frac{1}{i}(A \succ B \succ C)$, it follows that $\lim_{i\to\infty} p_i = p$ and for any i, $A \succ B \succ C \in f(p_i)$, but $A \succ B \succ C \notin f(p)$.

As we have noted before, there is generally a possibility of ties for PFs, and sometimes a PF is not defined for these cases (for example, we have not defined how they should be broken for STV). We can use the continuity property to gain some insight into how ties should be broken. For any $S \subseteq (\mathbb{Q}^{\geq 0})^{m!}$, let $C(S)$ be the *closure* of S, that is, $C(S)$ is the smallest set such that for any infinite sequence $p_1, p_2, \ldots$ in S, if $\lim_{i\to\infty} p_i = p$, then $p \in C(S)$. Let f_S be a PF that satisfies anonymity and homogeneity, defined over S. That is, $f_S : S \to 2^{L(A)} - \emptyset$. The *minimal continuous extension* of f_S is the PF $f_{C(S)} : C(S) \to 2^{L(A)} - \emptyset$ such that for any $p \in C(S)$ and any $r \in L(A)$, $r \in f_{C(S)}(p)$ if and only if there exists a sequence $p_1, p_2, \ldots$ in S such that $\lim_{i\to\infty} p_i = p$ and for any i, $r \in f_S(p_i)$. The following lemma will be useful in our study of STV.

Lemma 3 *Suppose we have two SRSFs f, f_S that have the same score function s, but f is defined over $(\mathbb{Q}^{\geq 0})^{m!}$, and f_S over a set $S \subseteq (\mathbb{Q}^{\geq 0})^{m!}$ such that $C(S) = (\mathbb{Q}^{\geq 0})^{m!}$. If for any $r \in L(A)$, there exists a profile p_r such that $f(p_r) = \{r\}$, then f is the minimal continuous extension of f_S.*

Proof: By Proposition 6, f is continuous. On the other hand, for any $p \in (\mathbb{Q}^{\geq 0})^{m!}$ with $r \in f(p)$, for any $i \in \mathbb{N}$, $f(p + \frac{1}{i}p_r) = \{r\}$. Because $C(S) = (\mathbb{Q}^{\geq 0})^{m!}$, for every $i \in \mathbb{N}$, there exists a point $p_i \in S$ sufficiently close to $p + \frac{1}{i}p_r$ such that $f(p_i) = \{r\}$, because s is continuous and at $p + \frac{1}{i}p_r$, for any $r' \in L(A)$ with $r \neq r'$, $s(p + \frac{1}{i}p_r, r) - s(p + \frac{1}{i}p_r, r') > 0$. So, $p_1, p_2, \ldots$ is a sequence of points in S with for any i, $r \in f_S(p_i)$; therefore any continuous extension must have $r \in f(p)$. ∎

7 Single Transferable Vote (STV)

In this section, we study the Single Transferable Vote (STV) PF in detail, for two reasons. First, it is a commonly used PF, so it is of interest in its own right. Second, it gives a good illustration of a number of subtle technical phenomena, and a precise understanding of these phenomena is likely to be helpful in the analysis of other PFs. We recall that under STV, in each round, the alternative that is ranked first (among the remaining alternatives) the fewest times is removed from all the votes and ranked the lowest among the remaining alternatives, that is, just above the previously removed alternative. We note that when an alternative is removed, all the votes that ranked it first *transfer* to the next remaining alternative in that vote. The number of votes ranking an alternative first is that alternative's *plurality score* in that round. One key issue is determining how ties in a round should be broken, that is, what to do if multiple alternatives have the lowest plurality score in a round. We will at first ignore this and show that STV is an ERSF. (This proof resembles the earlier Conitzer-Sandholm noise model but is much clearer in the language of scoring functions.)

Theorem 2 *When restricting attention to profiles without ties, STV is an ERSF.*

Proof: For $l \in L(A)$, let $l(i)$ be the ith-ranked alternative in l. Let $s_1(v, r) = 0$ if $r(m) = v(1)$, and $s_1(v, r) = 1$ otherwise. That is, a ranking receives a point for a vote if and only if the ranking does not rank the alternative ranked first in the vote last. Consider the alternative a with the lowest plurality score; the rankings that win under s_1 are exactly the rankings that rank a last. Now, let $s_2(v, r) = 0$ if *either* $r(m-1) = v(1)$, *or* $r(m) = v(1)$ and $r(m-1) = v(2)$; and $s_2(r, v) = 1$ otherwise. That is, a ranking receives a point for a vote *unless* the ranking ranks the first alternative in the vote second-to-last, or the ranking ranks the first alternative in the vote last and the second alternative in the vote second-to-last. If we look at rankings that survived s_2—the rankings that ranked the alternative a with the lowest plurality score last—a ranking that ranks b ($\neq a$) second-to-last will fail to receive a point for every vote that ranks b first, and for every vote that ranks a first and b second. That is, it fails to receive a point for every vote that ranks b first in the second iteration

of STV. Hence, the rankings that survive s_2 are the ones that rank the alternative that receives the fewest votes in the second iteration of STV second-to-last. More generally, let $s_k(v,r) = 0$ if, letting $b = r(m-k+1)$, for every a such that $v^{-1}(a) < v^{-1}(b)$, $r^{-1}(a) > r^{-1}(b) = m-k+1$; and $s_k(v,r) = 0$ otherwise. That is, a ranking receives a point for a vote *unless* the alternative b ranked kth-to last by r is preceded in v only by alternatives ranked after b in r. Given that r has not yet been eliminated and is hence consistent with STV so far, the latter condition holds if and only if b receives v's vote in the kth iteration of STV. ∎

In fact, we can break ties in STV simply according to the scoring functions used in the proof of Theorem 2. We will call the resulting PF *ERSF-STV*. ERSF-STV is an ERSF and hence consistent. By Theorem 1 and Proposition 3, this means that ERSF-STV is an MLE when there is an upper bound on the number of votes. Does there exist a tiebreaking rule for STV such that it is an SRSF, that is, so that it is an MLE without a bound on the number of votes? We will show that the answer is negative. To do so, we consider one particular tiebreaking rule. Under this rule, when multiple alternatives are tied to be eliminated, we have a choice of which one is eliminated. A ranking is among the winning rankings if and only if there is some sequence of such choices that results in this ranking. We call the resulting PF *parallel-universes tiebreaking STV (PUT-STV)*. (Every choice can be thought of as leading to a separate parallel universe in which STV is executed.)

Lemma 4 *PUT-STV is the minimal continuous extension of STV defined on non-tied profiles.*

Proof: Let f_{STV} be the STV PF restricted to the set S of non-tied profiles, and let $f_{PUT-STV}$ be PUT-STV. We first prove that for any tied profile $p = (t_1, \ldots, t_{m!})$ and any $r \in f_{PUT-STV}(p)$, there exists a sequence of points $p_1, p_2, \ldots \in S$ such that $\lim_{i\to\infty} p_i = p$ and for any i, $r \in f_{STV}(p_i)$. From this, it will follow that any continuous extension of f_{STV} must include all of the rankings that win under $f_{PUT-STV}$ among the winners. Let N be a positive integer such that for any $i \leq m!$, $Nt_i \in \mathbb{Z}$. Let $n = |Np|$, that is, $n = \sum_{i=1}^{m!} Nt_i$. For any $a \in A$ and any $r \in f_{PUT-STV}(p)$ such that $r = a_{i_1} \succ \ldots \succ a_{i_m}$, where $(i_1, \ldots, i_m)$ is a permutation of $(1, \ldots, m)$, let $v_{a,r} = a \succ a_{i_1} \succ others$ if $a \neq a_{i_1}$, and $v_{a,r} = a_{i_1} \succ others$ if $a = a_{i_1}$. (These are complete linear orders in which the order of the others does not matter.) We let $p_r = \sum_{j=0}^{m-1} 2^j \sum_{k<m-j} v_{a_{i_k},r}$. We now show that for any $\epsilon > 0$, $p + \epsilon p_r \in S$ and $f_{STV}(p + \epsilon p_r) = \{r\}$.

For any $A' \subseteq A$ and any profile p over A, let $p|_{A'}$ be the profile over A' obtained by removing all alternatives in $A - A'$ from p. For any $j \leq m$, let $A_j = \{a_{i_1}, \ldots, a_{i_{m-j}}\}$. For any profile p^*, subset of alternatives $A' \subseteq A$, and any alternative a, let $Pl(p^*|_{A'}, a)$ be the number of times that a is ranked first in the votes in $p^*|_{A'}$. We note that because $r \in f_{PUT-STV}(p)$, for any $j \leq m-1$, any $k < m-j$, $Pl(p|_{A_j}, a_{i_k}) \geq Pl(p|_{A_j}, a_{i_{m-j}})$. We have:

$$
\begin{aligned}
&Pl((p+\epsilon p_r)|_{A_j}, a_{i_k}) = Pl(p|_{A_j}, a_{i_k}) + \epsilon Pl(p_r|_{A_j}, a_{i_k}) \\
\geq &Pl(p|_{A_j}, a_{i_k}) + \epsilon 2^j > Pl(p|_{A_j}, a_{i_k}) + \epsilon \sum_{q=0}^{j-1} 2^q \\
= &Pl(p|_{A_j}, a_{i_k}) + \epsilon Pl(p_r|_{A_j}, a_{i_{m-j}}) \\
\geq &Pl(p|_{A_j}, a_{i_{m-j}}) + \epsilon Pl(p_r|_{A_j}, a_{i_{m-j}}) \\
= &Pl((p+\epsilon p_r)|_{A_j}, a_{i_{m-j}})
\end{aligned}
$$

Hence, for any $j \leq m-1$, in round j, $a_{i_{m-j}}$ is the alternative in A_j that is ranked first in the votes in $(p+\epsilon p_r)|_{A_j}$ (strictly) the fewest times. It follows that $f_{STV}(p+\epsilon p_r) = \{r\}$.

All that remains to show is that $f_{PUT-STV}$ is continuous, that is, for any sequence $p_1, p_2, \ldots \in S$ for which $\lim_{i\to\infty} p_i = p$ and there exists an $r \in L(A)$ such that for any i, $r \in f_{STV}(p_i)$, we have that $r \in f_{PUT-STV}(p)$. Again, let $r = a_{i_1} \succ \ldots \succ a_{i_m}$ and $A_j = \{a_{i_1}, \ldots, a_{i_{m-j}}\}$. Because for all i and $k < m-j$, $Pl(p_i|_{A_j}, a_{i_{m-j}}) < Pl(p_i|_{A_j}, a_{i_k})$, by the continuity of Pl, we

have $Pl(p|_{A_j}, a_{i_{m-j}}) \leq Pl(p|_{A_j}, a_{i_k})$. Hence, under PUT-STV, it is possible to eliminate $a_{i_{m-j}}$ in the $j+1$th round, completing the proof. ∎

Lemma 5 *PUT-STV is not consistent.*

Proof: Consider the following profile of votes, where A, B, and C are alternatives: $2(A \succ B \succ C) + 0(A \succ C \succ B) + 1(B \succ A \succ C) + 1(B \succ C \succ A) + 1(C \succ A \succ B) + 1(C \succ B \succ A)$. All alternatives are tied in the first round, and we split into three parallel universes. In the universe where A is eliminated, the $A \succ B \succ C$ votes transfer to B, and B is left as the only possible winner, producing the ranking $B \succ C \succ A$. In the universe where B is eliminated, the $B \succ A \succ C$ and $B \succ C \succ A$ votes transfer evenly to A and C, leaving us with another tie between A and C, and hence the rankings $A \succ C \succ B$ and $C \succ A \succ B$ are produced. Similarly, in the universe where C is eliminated first, the rankings $A \succ B \succ C$ and $B \succ A \succ C$ are produced. Ultimately, every ranking *except* $C \succ B \succ A$ is in the set of winning rankings.

By symmetry, under the profile $0(A \succ B \succ C) + 2(A \succ C \succ B) + 1(B \succ A \succ C) + 1(B \succ C \succ A) + 1(C \succ A \succ B) + 1(C \succ B \succ A)$, every ranking except $B \succ C \succ A$ wins. If we add the two profiles together, we obtain $2(A \succ B \succ C) + 2(A \succ C \succ B) + 2(B \succ A \succ C) + 2(B \succ C \succ A) + 2(C \succ A \succ B) + 2(C \succ B \succ A)$, which has all rankings in its output. But this violates consistency (which would require all rankings but $C \succ B \succ A$ and $B \succ C \succ A$ to win). ∎

Corollary 1 *PUT-STV is not an ERSF (and hence not an SRSF).*

Proof: This follows immediately from Proposition 4 and Lemma 5. ∎

This allows us to prove a property of STV in general:

Theorem 3 *STV is not an SRSF, even when restricting attention to non-tied profiles.*

Proof: Suppose that f_{STV} (restricted to the set S of non-tied profiles) is an SRSF defined by the score function s. By Lemma 4, $f_{PUT-STV}$ is the minimal continuous extension of f_{STV}. Also, for every $r \in L(A)$, it is easy to construct a (non-tied) profile p_r such that $f_{STV}(p_r) = \{r\}$. So, we can use Lemma 3 to conclude that $f_{PUT-STV}$ is the SRSF that results from using s on all profiles. However, by Corollary 1, we know that PUT-STV is not an SRSF, and we have the desired contradiction. ∎

Incidentally, PUT-STV is also computationally intractable (in a sense); we omit the proof due to space constraint. (We do not know if an analogous result holds for ERSF-STV.)

Theorem 4 *It is NP-complete to determine whether, given a profile p and an alternative a, one of the winning rankings under PUT-STV ranks a first.*

As it turns out, neither PUT-STV nor ERSF-STV corresponds to how ties are commonly broken under STV: rather, usually, if there is a tie, all of these alternatives are simultaneously eliminated. Mathematically, this leads to bizarre discontinuities; we omit further discussion due to space constraint.

8 Axiomatic characterization of SRSFs and ERSFs

Examining *social choice rules (SCRs)*, that is, functions that output one or more alternatives as the winner(s) (rather than one or more rankings), Young found the following axiomatic characterization of positional scoring functions [12]. (A similar characterization was given by Smith [11].)

He showed that all SCRs satisfying consistency, continuity, and neutrality—SCR analogues of the properties we considered—must be positional scoring functions, and all positional scoring functions satisfy these properties. Further, dropping continuity, he found that any consistent and neutral SCR must be equivalent to what in the language of this paper would be called an "extended" positional scoring function. These results lead to two natural analogous conjectures about PFs.

Conjecture 1 *Any PF that is consistent, continuous, and neutral must be an SRSF (and therefore an MLE).*

Conjecture 2 *Any PF that is consistent and neutral must be an ERSF (and therefore an MLE when the number of votes is bounded).*

It does not appear that these conjectures can be easily proven using Smith and Young's techniques.

9 Conclusions

The maximum likelihood approach provides a natural way for choosing a PF in settings where it makes sense to think there is a "correct" ranking. In this paper, we gave a characterization of the neutral MLE PFs, showing they coincide with the neutral SRSFs. We also considered ERSFs as a slight generalization and showed that for bounded numbers of votes they coincide with SRSFs. We considered key properties such as continuity and consistency, and gave several examples of SRSFs and ERSFs. We studied STV in detail, showing that it is an ERSF but not an SRSF, and discussed the implications for breaking ties under STV. Finally, we left some open questions concerning the complexity of ERSF tiebreaking for STV and whether consistency can be used to characterize the class of SRSFs/ERSFs.

We believe that these results will greatly facilitate the use of the maximum likelihood approach in (computational) social choice. Similar results can be obtained for social choice settings other than PFs—for example, for social choice rules that only choose the winning alternative(s), or for settings in which the inputs are not linear orders (but rather, for example, labelings of the alternatives as "approved" or "not approved", or partial orders, *etc.*).

Acknowledgements

We thank Felix Brandt, Zheng Li, Ariel Procaccia, and Bill Zwicker for very helpful discussions, and the reviewers for very helpful comments. The authors are supported by an Alfred P. Sloan Research Fellowship, a Johnson Research Assistantship, and a James B. Duke Fellowship, respectively; they are also supported by the NSF under award number IIS-0812113.

References

[1] John Bartholdi, III, Craig Tovey, and Michael Trick. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare*, 6:157–165, 1989.

[2] Vincent Conitzer, Andrew Davenport, and Jayant Kalagnanam. Improved bounds for computing Kemeny rankings. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 620–626, Boston, MA, 2006.

[3] Vincent Conitzer and Tuomas Sandholm. Common voting rules as maximum likelihood estimators. In *Proceedings of the 21st Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 145–152, Edinburgh, UK, 2005.

[4] Marie Jean Antoine Nicolas de Caritat (Marquis de Condorcet). Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. 1785. Paris: L'Imprimerie Royale.

[5] Mohamed Drissi and Michel Truchon. Maximum likelihood approach to vote aggregation with variable probabilities. Technical Report 0211, Departement d'economique, Universite Laval, 2002.

[6] Michael Fligner and Joseph Verducci. Distance based ranking models. *Journal of the Royal Statistical Society B*, 48:359–369, 1986.

[7] John Kemeny. Mathematics without numbers. In *Daedalus*, volume 88, pages 571–591. 1959.

[8] Guy Lebanon and John Lafferty. Cranking: Combining rankings using conditional models on permutations. In *International Conference on Machine Learning (ICML)*, pages 363–370, Sydney, Australia, 2002.

[9] Marina Meila, Kapil Phadnis, Arthur Patterson, and Jeff Bilmes. Consensus ranking under the exponential model. In *Proceedings of the 23rd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, Vancouver, BC, Canada, 2007.

[10] Roger B. Myerson. Axiomatic derivation of scoring rules without the ordering assumption. *Social Choice and Welfare*, 12(1):59–74, 1995.

[11] John H. Smith. Aggregation of preferences with variable electorate. *Econometrica*, 41(6):1027–1041, November 1973.

[12] H. Peyton Young. Social choice scoring functions. *SIAM Journal of Applied Mathematics*, 28(4):824–838, 1975.

[13] H. Peyton Young. Optimal voting rules. *Journal of Economic Perspectives*, 9(1):51–64, 1995.

[14] H. Peyton Young and Arthur Levenglick. A consistent extension of Condorcet's election principle. *SIAM Journal of Applied Mathematics*, 35(2):285–300, 1978.

[15] William S. Zwicker. Consistency without neutrality in voting rules: When is a vote an average? *Mathematical and Computer Modelling*, 2008.

Vincent Conitzer
Departments of Computer Science and Economics
Duke University
Durham, NC 27708, USA
Email: conitzer@cs.duke.edu

Matthew Rognlie
Duke University
Durham, NC 27708, USA
Email: matthew.rognlie@duke.edu

Lirong Xia
Department of Computer Science
Duke University
Durham, NC 27708, USA
Email: lxia@cs.duke.edu

Computing Spanning Trees in a Social Choice Context

Andreas Darmann, Christian Klamler and Ulrich Pferschy

Abstract

This paper combines social choice theory with discrete optimization. We assume that individuals have preferences over edges of a graph that need to be aggregated. The goal is to find a socially "best" spanning tree in the graph. As ranking all spanning trees is becoming infeasible even for small numbers of vertices and/or edges of a graph, our interest lies in finding algorithms that determine a socially "best" spanning tree in a simple manner. This problem is closely related to the minimum (or maximum) spanning tree problem in combinatorial optimization. Our main result shows that for the various underlying ranking rules on the set of spanning trees discussed in this paper the sets of "best" spanning trees coincide. Moreover, a greedy algorithm based on a transitive group ranking on the set of edges will always provide such a "best" spanning tree.

1 Introduction

In this paper we want to apply tools from social choice theory to topics from discrete optimization. Although these topics are historically separated, in recent years there have started attempts to combine these approaches. This is especially of interest whenever mathematical concepts (such as graphs) are used in problems where group decisions need to be made.

As an actual example one could think of a small village that has to install a water network or countries that need to agree on oil pipelines. Every homeowner in the village needs to be connected, however, there are many different ways to hook them up. A mathematical representation of such a situation could be done by a spanning tree on a graph that connects each pair of homeowners (i.e. vertices) in the village. However, different homeowners might have different preferences over the possible connections between the homeowners (i.e. edges of the graph). E.g. one homeowner might rather want to have it pass through his own garden than through a nice park, whereas another one might think the other way round.

Such aggregations of individual preferences are the major focus in social choice theory and many different aggregation rules do exist and have been studied and compared in the literature (see e.g. [8] or [10]). Our approach, however, will not analyse the aggregation of such individual preferences, but start off with a group ranking of the possible edges. This ranking of the edges does not necessarily allocate numerical values to the edges. As each spanning tree is a subset of the set of edges having a certain structure, we will - given the group ranking - try to rank the different spanning trees. This lies in the spirit of previous results on ranking sets of objects (for an overview see [2]). Such rankings could be based on Borda counts, on simple majority rule, be of lexicographic nature, etc; as for the edges, the ranking of the spanning trees does not have to be based on numerical values of the edges and does not need to assign a number to each tree. The goal for this type of problems is to find the spanning tree which is "best" w.r.t. such a relation on spanning trees. Ranking all spanning trees is, however, a difficult problem even for small number of vertices and/or edges given the quickly increasing number of feasible spanning trees. Our interest therefore lies in finding algorithms that determine a "best" spanning tree in a simple manner. This problem is closely related to the minimum (or maximum) spanning tree problem, a classical problem

in discrete optimization. The minimum spanning tree problem has numerous applications in various fields and can be solved efficiently by greedy algorithms [1].

The main result in this paper shows that irrespective of the underlying ranking rules discussed in this paper the set of "best" spanning trees coincide. What's more, using a greedy algorithm to determine a maximal spanning tree based on a transitive group ranking on the set of edges will always provide a "best" spanning tree.

2 Formal framework

2.1 Preliminaries

Let $G = (V, E)$ be an undirected graph where V denotes the set of nodes and E denotes the set of edges. Let $n := |V|$ and $m := |E|$. A subset $T \subseteq E$ is called spanning tree of G, if the subgraph (V, T) of G is acyclic and connected. Let τ denote the set of spanning trees of G. Let $I = \{1, ..., k\}$ denote a finite set of individuals. For every individual i, $1 \leq i \leq k$, the preference order P_i on E is assumed to be a linear order on E (i.e. P_i is assumed to be complete, transitive and asymmetric). A k-tuple $\pi = (P_1, P_2, ..., P_k)$ is called a preference profile and $\wp$ denotes the set of admissible preference profiles. A complete order $\succsim$ consists of an asymmetric part $\succ$ and a symmetric part $\sim$ respectively. Given a complete order $\succsim_\kappa$ on τ, we call a tree T *best tree with respect to* $\succsim_\kappa$ if there is no $B \in \tau$ with $B \succ_\kappa T$.

Remark. Let $T_1, T_2 \in \tau$. Having assigned a real number $w(e)$ to each $e \in E$, the minimum spanning tree problem is the problem of finding a best tree with respect to the relation $T_1 \succsim T_2 :\Longleftrightarrow \sum_{e \in T_1} w(e) \leq \sum_{e \in T_2} w(e)$. Analogously, a maximum spanning tree is a best tree with respect to the relation $T_1 \succsim T_2 :\Longleftrightarrow \sum_{e \in T_1} w(e) \geq \sum_{e \in T_2} w(e)$.

We first present basic complete orders on the set E of edges from which orders on the set τ of spanning trees of G are derived.

2.2 Basic orders on E

Definition 2.1 *Given P_i, individual i's Borda count of an edge $e \in E$ is given by $B_i(e) := |\{f \in E : eP_if\}|$. The total Borda count of edge e is defined by $B(e) := \sum_{i \in I} B_i(e)$. For $e, f \in E$ we define $e \succsim_b f :\Longleftrightarrow B(e) \geq B(f)$.*

Definition 2.2 *Let $e, f \in E$. Then we define the Simple Majority-order on E by $e \succsim_{sm} f :\Longleftrightarrow |\{i \in I : eP_if\}| \geq |\{i \in I : fP_ie\}|$.*

For all $i \in I$ we define the singleton set S_i^t representing individual i's top choice by $S_i^t := \{e \in E | eP_if \, \forall f \in E \setminus \{e\}\}$. Furthermore let, for all $i \in I$, the set E be partitioned into a set $S_i \subset E$ of edges individual i approves of and a set $E \setminus S_i$ individual i disapproves of.

Definition 2.3 *Let $e, f \in E$. The Approval count of e is defined by $A(e) := |\{i \in I : e \in S_i\}|$. The Approval-order $\succsim_a$ is then defined by $e \succsim_a f :\Longleftrightarrow A(e) \geq A(f)$.*

Definition 2.4 *Let $e, f \in E$. The Plurality count of e is $Pl(e) := |\{i \in I : e \in S_i^t\}|$. The Plurality-order $\succsim_{pl}$ is defined by $e \succsim_{pl} f :\Longleftrightarrow Pl(e) \geq Pl(f)$.*

The relations $\succsim_b$, $\succsim_a$ and $\succsim_{pl}$ are *weak orders* on E, i.e. these relations are complete and transitive. The relation $\succsim_{sm}$ is a complete order as well, but in general $\succsim_{sm}$ is not transitive and hence not a weak order. Thus, in $\succsim_{sm}$ preference cycles may occur. To overcome this

inconvenience, one might be interested in procedures that transform a complete but not transitive order into a weak order. We introduce Copeland's procedure [7], other possibilities are e.g. Slater's procedure [11] or Black's procedure [5].

Definition 2.5 *Let $\succsim_n$ be a complete order on E and let $e, f \in E$. Let*

$$s(e,f) := \begin{cases} 1 & \text{if } e \succ_n f \\ 0 & \text{if } e \sim_n f \\ -1 & \text{if } e \prec_n f \end{cases}$$

be the score of e versus f. Let $z(e) := \sum_{g \in E} s(e,g)$. Then we define $e \succsim_{cl} f :\Longleftrightarrow z(e) \geq z(f)$ and call $\succsim_{cl}$ Copeland's order on E.

Remark. Note that $s(e,f) = -s(f,e)$ holds for all $e, f \in E$.

Obviously Copeland's order is a weak order on E. Thus setting $\succsim_n := \succsim_{sm}$ and determining the corresponding Copeland's order for example yields a weak order on E based on the Simple Majority-order.

3 Some complete orders on τ

We first present three weak orders on τ that are based on weak orders on E presented in Section 2.2. The Borda-order introduced in the following definition ranks $T_1 \in \tau$ not lower than $T_2 \in \tau$, if the sum of Borda counts of the edges contained in T_1 is at least as high as the sum of Borda counts of the edges of T_2.

Definition 3.1 *For $T \in \tau$ we define the Borda count of T by $B(T) := \sum_{e \in T} B(e)$. Then the Borda-order $\succsim_B$ on τ is defined by letting, for all $T_1, T_2 \in \tau$,*

$$T_1 \succsim_B T_2 :\Longleftrightarrow B(T_1) \geq B(T_2) \ .$$

Analogously, a best tree with respect to the Approval-order (Plurality-order) on τ is a tree maximizing the edge-sum of Approval counts (Plurality counts).

Definition 3.2 *For $T \in \tau$ the Approval count of T is defined by $A(T) := \sum_{e \in T} A(e)$. The Approval-order $\succsim_A$ on τ is defined by letting, for all $T_1, T_2 \in \tau$,*

$$T_1 \succsim_A T_2 :\Longleftrightarrow A(T_1) \geq A(T_2) \ .$$

Definition 3.3 *For $T \in \tau$ the Plurality count of T is defined by $Pl(T) := \sum_{e \in T} Pl(e)$. Then we define the Plurality-order $\succsim_P$ on τ by letting, for all $T_1, T_2 \in \tau$,*

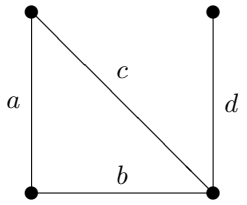
$$T_1 \succsim_P T_2 :\Longleftrightarrow Pl(T_1) \geq Pl(T_2) \ .$$

Having assigned Borda counts (Approval counts, Plurality counts) to the edges $e \in E$, a best tree with respect to the Borda-order (Approval-order, Plurality-order) is a tree with maximum Borda count (Approval count, Plurality count). This approach can be generalized as follows.

Definition 3.4 *Let τ be the set of spanning trees of G and let $\succsim$ be a weak order on E. A tree $M \in \tau$ is called max-spanning tree if and only if for every edge $f = \{i,j\}$, $f \notin M$,*

$$f \precsim e$$

holds for all $e \in M$ that are part of the unique simple path between i and j in M.



1	2	3
a	b	c
b	c	a
c	a	b
d	d	d

Figure 1: An undirected graph $G = (V, E)$ and a preference profile for three voters.

Remark. The above definition of a max-spanning tree is a generalization of the path optimality condition for the maximum spanning tree problem stated in [1]. Note that for Definition 3.4 $\succsim$ does not need to be based on numerical values; that is, there does not have to be a number assigned to each edge.

A tree with maximum Borda count (Approval count, Plurality count) and, a max-spanning tree in general, can be determined efficiently by applying greedy algorithms – such as Prim's or Kruskal's algorithm (for details see [1]) – for the maximum spanning tree problem. For example, a "generalized" version of Kruskal's algorithm to compute a max-spanning tree works as follows:
Arrange the edges $e \in E$ in non-increasing order according to $\succsim$ and iteratively add an edge to the solution set X (which is empty at the beginning) such that socially preferred edges are taken first. I.e. first add to X an edge that no other edge is socially preferred to, continue with the "next best" edge, etc. If adding an edge creates a cycle, the edge simply is ignored and we go on with the next edge.

As mentioned above, the Simple Majority-order $\succsim_{sm}$ on E however is not a weak order because in general it is not transitive, and thus preference cycles may occur. Hence the Simple Majority-order does not seem to immediately indicate an order on τ analogous to the Borda, Approval or Plurality case. Nevertheless the complete order $\succsim_{sm}$ on E can be used to compare two trees. The idea of the following concept for comparing two trees based on a given (not necessarily transitive) complete order $\succsim$ on E is to remove the edges the trees have in common and to pairwise compare the remaining edges according to $\succsim$.

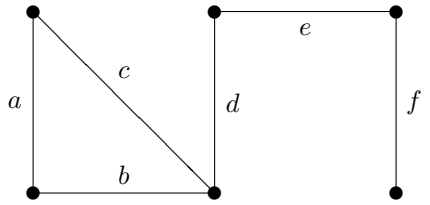
Definition 3.5 *Let $\succsim$ be a complete order on E and let $T_1, T_2 \in \tau$. Furthermore, let $\tilde{T}_1 := T_1 \setminus T_2$ and $\tilde{T}_2 := T_2 \setminus T_1$. Then we define*

$$T_1 \succsim_S T_2 \ :\Longleftrightarrow \sum_{a \in \tilde{T}_1} \sum_{b \in \tilde{T}_2} s(a,b) \geq 0 \ ,$$

where, for $a, b \in E$, $s(a,b)$ denotes the score of a versus b.

Remark. $\succsim_S$ is a complete order on τ. Furthermore it is worth mentioning that in Definition 3.5, as well as in Definition 3.4, $\succsim$ does not need to be of numerical nature, i.e. $\succsim$ does not have to allocate numbers to the edges.

Example 1 *Let $\succsim = \succsim_{sm}$. For the graph displayed in Figure 1 there exist three spanning trees: $T_1 := \{a, b, d\}$, $T_2 := \{b, c, d\}$ and $T_3 := \{a, c, d\}$. Given the preference profile in Figure 1 we get $a \succ_{sm} b$ because we have aP_1b and aP_3b, whereas only voter 2 prefers b to*



1	2	3
c	f	b
a	e	f
b	d	a
d	c	d
e	a	e
f	b	c

Figure 2: On the left side an undirected graph $G = (V, E)$ and on the right side a corresponding preference profile for three voters are displayed.

a. Analogously we get $b \succ_{sm} c$ and $c \succ_{sm} a$.
Because of $T_1 \setminus T_2 = \{a\}$ and $T_2 \setminus T_1 = \{c\}$ we get $T_2 \succ_S T_1$ due to $c \succ_{sm} a$. $T_1 \setminus T_3 = \{b\}$ and $T_3 \setminus T_1 = \{c\}$ yield $T_1 \succ_S T_3$ because of $b \succ_{sm} c$. Finally, we get $T_3 \succ_S T_2$ because $T_2 \setminus T_3 = \{b\}$ and $T_3 \setminus T_2 = \{a\}$ hold and $a \succ_{sm} b$ is satisfied. Thus we have $T_3 \succ_S T_2$, $T_2 \succ_S T_1$ and $T_1 \succ_S T_3$. Hence a best tree with respect to $\succsim_S$ does not exist in this example.

As the previous example shows, if $\succsim$ is not transitive (e.g. if $\succsim = \succsim_{sm}$) a best tree with respect to $\succsim_S$ in general does not exist because preference cycles may occur. Furthermore, as the following example shows, a tree T_g output by a greedy algorithm (a "generalized" version of Kruskal's algorithm) that is based on arranging edges e according to $|\{f \in E : e \succ f\}|$ in non-increasing order might be the *worst* tree according to $\succsim_S$, i.e. every other spanning tree T satisfies $T \succ_S T_g$.

Example 2 *Given the graph and the preference profile shown in Figure 2, the Simple Majority-order on $E = \{a, b, c, d, e, f\}$ is of the following form:*

$a \succ_{sm} b$	$b \succ_{sm} d$	$c \succ_{sm} a$	$d \succ_{sm} e$	$f \succ_{sm} c$
$a \succ_{sm} d$	$b \succ_{sm} e$	$c \succ_{sm} b$	$e \succ_{sm} c$	$f \succ_{sm} d$
$a \succ_{sm} e$	$b \succ_{sm} f$	$d \succ_{sm} c$	$f \succ_{sm} a$	$f \succ_{sm} e$

Thus, the edge f is superior to four edges according to the Simple Majority-order, the edges a and b are superior to three edges, c and d to two edges and e to one edge only. It is easy to see that a generalized version of Kruskal's algorithm that arranges the edges according to the Simple Majority-wins outputs the tree $T_g = \{f, a, b, d, e\}$. Altogether the graph contains three spanning trees, the two others being $T_1 = \{b, c, d, e, f\}$ and $T_2 = \{a, c, d, e, f\}$. We get $T_1 \setminus T_g = \{c\}$ and $T_g \setminus T_1 = \{a\}$, implying $T_1 \succ_S T_g$. Furthermore we have $T_2 \setminus T_g = \{c\}$ and $T_g \setminus T_2 = \{b\}$ and hence $T_2 \succ_S T_g$. I.e. according to $\succsim_S$ every other spanning tree of the graph is strictly preferred to T_g.

In the next section however we show that a best tree with respect to $\succsim_S$ always exists and can be computed efficiently if $\succsim$ is assumed to be a weak order on E. Thus it seems to be a reasonable approach to use Copeland's order (Slater's order, Black's order) to establish from $\succsim_{sm}$ a weak order on the set E; Copeland's order (Slater's order, Black's order) then might be used in order to determine $\succsim_S$.

4 Comparing trees

In the previous section we presented three weak orders on E that yield weak orders on τ in an intuitive way. Further we presented a complete order $\succsim_S$ on τ that consists of pairwise

comparisons of edges based on a complete order $\succsim$ on E. If $\succsim$ is not transitive (e.g. relation $\succsim_{sm}$), preference cycles may occur and a best tree with regard to $\succsim_S$ might not exist. To overcome this difficulty, we presented methods to transform a complete order on E that is not transitive into a transitive order on E.

In this Section we will show that assuming $\succsim$ to be a weak order implies the existence of a best tree with respect to $\succsim_S$. Moreover, we will show the equivalence of four orders on τ that are based on a weak order $\succsim$ in the sense that a best tree with respect to one order is always a best tree with respect to each of the other orders. This implies that, for all four orders, a best tree with respect to the regarded order always exists and can be computed efficiently.

In what follows, $\succsim$ is assumed to be a given weak order on E. Note that $\succsim$ is not necessarily based on numerical values assigned to the edges.

4.1 Three more complete orders on τ

Given $T_1, T_2 \in \tau$, we use the notation $\tilde{T}_1 := T_1 \setminus T_2$, $\tilde{T}_2 := T_2 \setminus T_1$ and $r := |\tilde{T}_1|$ within Section 4.1 for convenience.

Based on the weak order $\succsim$ on E the three complete orders $\succsim_{lex}$, $\succsim_{mxn}$, and $\succsim_{ps}$ on the set τ of spanning trees of G are defined. The following maxmin-order on τ is derived from the maxmin-order on sets presented in [2]. According to the maxmin-order, a spanning tree T_1 is preferred to a spanning tree T_2 if either a "best" edge in $T_1 \setminus T_2$ is preferred to a socially most attractive edge in $T_2 \setminus T_1$ or, in case of indifference between these two edges, if the least accepted edge in $T_1 \setminus T_2$ is preferred to the socially worst edge in $T_2 \setminus T_1$.

Definition 4.1 *Let $T_1, T_2 \in \tau$. Then we define the maxmin-order $\succsim_{mxn}$ on τ by*

$$\begin{aligned} T_1 \succsim_{mxn} T_2 \ :\Longleftrightarrow \ & [\tilde{T}_1 = \emptyset \text{ or} \\ & \max \tilde{T}_1 \succ \max \tilde{T}_2 \text{ or} \\ & (\max \tilde{T}_1 \sim \max \tilde{T}_2 \text{ and } \min \tilde{T}_1 \succsim \min \tilde{T}_2)] \end{aligned}$$

Analogously we define the leximax order on trees based on the leximax order on sets presented in [2].

Definition 4.2 *Let $T_1, T_2 \in \tau$. Further let $\tilde{T}_1 := \{e_1, e_2, ..., e_r\}$, $\tilde{T}_2 := \{f_1, f_2, ..., f_r\}$ such that $e_i \succsim e_{i+1}$ and $f_i \succsim f_{i+1}$ holds for $1 \leq i \leq r-1$. Then the leximax order $\succsim_{lex}$ on τ is defined by*

$$\begin{aligned} T_1 \succsim_{lex} T_2 \ :\Longleftrightarrow \ & [\tilde{T}_1 = \emptyset \text{ or} \\ & e_i \sim f_i \text{ for all } 1 \leq i \leq r \text{ or} \\ & (\exists j \in \{1, ..., r\} \text{ such that} \\ & e_i \sim f_i \text{ for all } i < j \text{ and } e_j \succ f_j)] \end{aligned}$$

A third approach is to rank the edges of the disjoint union of the two trees according to $\succsim$. For the resulting ranking a positional scoring concept is used to compare the two regarded trees. This approach adapts the concept of the positional scoring procedures presented in [6].

Definition 4.3 *Let $T_1, T_2 \in \tau$. Further let $\tilde{T}_1 \cup \tilde{T}_2 := \{d_1, d_2, ...d_{2r}\}$ such that $d_i \succsim d_{i+1}$ holds for $1 \leq i \leq 2r-1$. Let $b : E \to \mathbb{R}$ be strictly increasing according to $\succsim$, that is, for $1 \leq i \leq 2r-1$, $b(d_i) = b(d_{i+1})$ if $d_i \sim d_{i+1}$ and $b(d_i) > b(d_{i+1})$ if $d_i \succ d_{i+1}$.*
Let $b(\tilde{T}_1) := \sum_{e \in \tilde{T}_1} b(e)$ and $b(\tilde{T}_2) := \sum_{f \in \tilde{T}_2} b(f)$.

Then we define

$$T_1 \succsim_{ps} T_2 \ :\Longleftrightarrow b(\tilde{T}_1) \geq b(\tilde{T}_2) \ .$$

Remark. The edges being ranked as in the above definition, and having assigned $b(d_j)$ scoring points to the edge in the j-th position as in Definition 4.3, a best tree with respect to $\succsim_{ps}$ is a spanning tree that maximizes the sum of scoring points.

4.2 Results

This Section is organized as follows.

Theorem 4.1 states that a max-spanning tree can be computed in polynomial time[1]. We afterwards show that a max-spanning tree (as defined in Definition 3.4) corresponds to a best tree with respect to $\succsim_S$ and to a best tree with respect to each of the orderings on τ defined in Section 4.1. Vice versa, a best tree with respect to one of these orderings always corresponds to a max-spanning tree. We summarize these results in Theorem 4.4. Together with Theorem 4.1, Theorem 4.4 implies that a best tree with respect to each of these orderings can be determined in polynomial time.

Theorem 4.1 *A max-spanning tree can be computed in $\mathcal{O}(m + n \log n)$ time.*

Proof. The Theorem immediately follows from the fact that the maximum spanning tree problem can be solved in $\mathcal{O}(m + n \log n)$ time [1]. □

Theorem 4.2 *A tree $M \in \tau$ is a max-spanning tree if and only if there is no tree $B \in \tau$ such that*

$$B \succ_{lex} M$$

holds.

Proof.
"⇒": Assume there exists $B \in \tau$ with $B \succ_{lex} M$. Let $B \setminus M := \{f_1, ..., f_r\}$ and $M \setminus B := \{e_1, ..., e_r\}$ for some $r \geq 1$ such that

$$f_i \succsim f_{i+1} \text{ and } e_i \succsim e_{i+1} \text{ for all } 1 \leq i \leq r-1 \ . \tag{1}$$

Obviously $B \succ_{lex} M$ implies $f_1 \succsim e_1$. We now show that $f_1 \sim e_1$ must hold.

- Assume that $f_1 \succ e_1$ holds. Because M is a tree adding f_1 to M yields a cycle K_1. Since B is a tree not all edges of K_1 can be contained in B and hence K_1 must contain an edge $\tilde{e} \in \{e_1, e_2, ..., e_r\}$. This means that there is an edge e in K_1 that satisfies $f_1 \succ \tilde{e}$ because both $f_1 \succ e_1$ and $e_1 \succsim \tilde{e}$ hold. This is a contradiction to the fact that M is a max-spanning tree.

Thus we have $f_1 \sim e_1$ and by our assumption there exists an index $1 < j \leq r$ such that

$$e_i \sim f_i \tag{2}$$

for all $i < j$ and

$$e_j \prec f_j \tag{3}$$

hold. Note that because of (1) and (2)

$$f_i \succsim e_{i+1} \tag{4}$$

[1] Note that, in case $P \neq NP$, this does not have to hold if the order $\succsim$ on E is not assumed to be given but needs to be determined from the individual preferences. For example, computing $\succsim$ using Kemeny's or Dodgson's rule is NP-hard [3]. Thus, the whole process of computing $\succsim$ according to Kemeny's rule and determining a max-spanning tree thereafter would be NP-hard.

holds for all $i < j-1$.
Adding f_{j-1} to M creates a cycle K. Because M is a max-spanning tree inequality

$$e \succsim f_{j-1} \tag{5}$$

holds for all $e \in K$. Now add K to B. Removing f_{j-1} from $B \cup K$ yields a connected graph $A := (B \cup K) \setminus \{f_{j-1}\}$ because f_{j-1} is part of K. Because B is a tree, for every cycle C contained in A an edge $e \in K$ (recall that $e \in M$ holds) must be part of C. Hence removing such edges e from A until we get an acyclic graph yields a tree A_1 which must be of the form $A_1 = B \setminus \{f_{j-1}\} \cup \{e\}$ for some $e \in M$. Note that due to (1) and (3) $f_{j-1} \succsim f_j \succ e_j$ holds. Hence (5) implies that

$$e \in \{e_1, e_2, ..., e_{j-1}\} \tag{6}$$

must hold. Now we show that $A_1 \succ_{lex} M$ must be satisfied:

- Case (i): $e = e_1$. Then $M \setminus A_1 = \{e_2, e_3, ..., e_{j-1}, e_j, ..., e_r\}$ and $A_1 \setminus M = \{f_1, f_2, ..., f_{j-2}, f_j, ..., f_r\}$. (4) states $f_i \succsim e_{i+1}$ for all $i < j-1$ and (3) states $f_j \succ e_j$ which implies $A_1 \succ_{lex} M$.
- Case (ii): $e = e_{j-1}$. Then we get $M \setminus A_1 = \{e_1, ..., e_{j-2}, e_j, ..., e_r\}$ and $A_1 \setminus M = (f_1, ..., f_{j-2}, f_j, ..., f_r)$. Obviously $A_1 \succ_{lex} M$ holds in this case since (2) and (3) hold.
- Case (iii): $e = e_k$ for some $1 < k < j-1$. In this case we have

  $$M \setminus A_1 = \{e_1, ..., e_{k-1}, e_{k+1}, ..., e_{j-1}, e_j, ..., e_r\}$$

  and $A_1 \setminus M = (f_1, ..., f_{k-1}, f_k, ..., f_{j-2}, f_j, ..., f_r)$. Again because of (2), (4) and (3) we get $A_1 \succ_{lex} M$.

In all three cases we have $A_1 \succ_{lex} M$ with $A_1 \setminus M = \{f_1, f_2, ..., f_{j-2}, f_j, ..., f_r\}$. I.e. given $B \succ_{lex} M$ we can create a tree $A_1 \succ_{lex} M$ such that $A_1 \setminus M$ equals $B \setminus M$ without edge f_{j-1}.
Repeating this procedure $j-2$ times yields a tree $A_{j-1} \succ_{lex} M$ with $A_{j-1} \setminus M = \{f_j, ..., f_r\}$. Note that $M \setminus A_{j-1} = \{e_j, ..., e_r\}$ must hold due to (6). Now recall that $f_j \succ e_j$ was stated in (3). Furthermore recall that at the beginning of this proof we showed that $f_1 \succ e_1$ does not hold. Therewith it is proven that there cannot exist a tree $C \succ_{lex} M$ with $C \setminus M = \{c_1, ..., c_p\}$, $M \setminus C = \{d_1, ..., d_p\}$ for some $p \geq 1$, where $c_i \succsim c_{i+1}$ and $d_i \succsim d_{i+1}$ hold for all $1 \leq i \leq p-1$, such that $c_1 \succ d_1$ holds. This contradicts to the existence of A_{j-1}.
"$\Leftarrow$": Assume M is not a max-spanning tree. This assumption implies the existence of an edge $e \in E \setminus M$ such that the unique cycle K in $M \cup \{e\}$ contains an edge f such that $f \prec e$ holds. Now consider the tree $T := M \cup \{e\} \setminus \{f\}$. We get $T \setminus M = \{e\}$ and $M \setminus T = \{f\}$. Thus we get $T \succ_{lex} M$ which is a contradiction. □

Remark. It should be mentioned that Bern and Eppstein [4] state without proof the observation that a minimum spanning tree lexicographically minimizes the vector of edge lengths. Assigning a real number $w(e)$ to each $e \in E$ and setting $e \succsim f :\Longleftrightarrow w(e) \leq w(f)$ for $e, f \in E$, where $\leq$ denotes the common less-than-or-equal relation on $\mathbb{R}$, this observation immediately follows from Theorem 4.2.

Theorem 4.3 *A tree $M \in \tau$ is a max-spanning tree if and only if there is no tree $B \in \tau$ such that $B \succ_S M$ holds.*

Proof.
"⇒": Assume the set $\beta := \{B \in \tau : B \succ_S M\}$ is not empty and let $E_B := B \cap M$ for all $B \in \beta$. Clearly, there exists a tree $B_1 \in \beta$ such that $|E_{B_1}| \geq |E_B|$ holds for all $B \in \beta$. Since every spanning tree contains exactly $n-1$ edges, $|E_{B_1}| = n-1$ means $B_1 = M$ in contradiction to $B_1 \in \beta$. Hence $0 \leq |E_{B_1}| \leq n-2$ holds. This implies that, for some $l \in \{1, 2, ..., n-1\}$ and some $e_i, f_i \in E$, $1 \leq i \leq l$, we have

$$\tilde{B_1} := B_1 \setminus M = \{f_1, f_2, ..., f_l\} \text{ with } f_1 \succeq f_2 \succeq ... \succeq f_l \tag{7}$$

and

$$M_1 := M \setminus B_1 = \{e_1, e_2, ..., e_l\} \text{ with } e_1 \succeq e_2 \succeq ... \succeq e_l \ . \tag{8}$$

Note that $f_1 \preceq e_1$ must hold because of Theorem 4.2. Hence we get

$$f_i \preceq e_1 \text{ for all } 1 \leq i \leq l \ . \tag{9}$$

Adding e_1 to B_1 yields a cycle K_1. Obviously not all edges of K_1 can be contained in M and thus K_1 contains at least one edge f_j, $1 \leq j \leq l$. This means that $A := B_1 \cup \{e_1\} \setminus \{f_j\}$ is a spanning tree of G. Now we show that $A \succ_S M$:

- We have $\tilde{A} := A \setminus M = \{f_1, f_2, ..., f_{j-1}, f_{j+1}, ..., f_l\}$ and $\tilde{M} := M \setminus A = \{e_2, e_3, ..., e_l\}$ for some $1 \leq j \leq l$. Hence we get

$$\begin{aligned} \sum_{f \in \tilde{A}} \sum_{e \in \tilde{M}} s(f,e) \quad &= \quad \sum_{f \in \tilde{B}_1} \sum_{e \in M_1} s(f,e) \\ &\quad - \sum_{e \in M_1 \setminus \{e_1\}} s(f_j, e) \ - \sum_{f \in \tilde{B}_1 \setminus \{f_j\}} s(f, e_1) - s(f_j, e_1) \ . \end{aligned}$$

  Note that $\sum_{f \in \tilde{B}_1} \sum_{e \in M_1} s(f,e) > 0$ holds because of $B_1 \succ_S M$ by definition of β. What's more, we have $s(f_j, e_1) \leq 0$ due to (9). It remains to show that

$$\sum_{f \in \tilde{B}_1 \setminus \{f_j\}} s(e_1, f) - \sum_{e \in M_1 \setminus \{e_1\}} s(f_j, e) \geq 0 \tag{10}$$

  holds.
  Theorem 4.2 yields $B_1 \precsim_{lex} M$. Obviously $B_1 \sim_{lex} M$ yields $B_1 \sim_S M$ in contradiction to the definition of β and thus $B_1 \prec_{lex} M$ must hold. Hence there exists a $1 \leq k \leq l$ such that $e_i \sim f_i$ holds for $i < k$ and $e_k \succ f_k$ is satisfied.
  If $e_1 \succ f_1$ then $\sum_{f \in \tilde{B}_1 \setminus \{f_j\}} s(e_1, f) = l - 1$ and (10) is satisfied since $\sum_{e \in M_1 \setminus \{e_1\}} s(f_j, e) \leq l-1$. Hence we assume $k \geq 2$ (note that $l = 1$ implies $k = 1$ and thus w.l.o.g. we assume $l \geq 2$ as well). Since $k \geq 2$ there exists an index r, $1 \leq r \leq k-1$, such that both $e_r \sim f_r$ and $e_r \succ f_{r+1}$ is satisfied, because $e_1 \succsim e_2 \succsim ... \succsim e_l$ holds (see (8)). Let q, $1 \leq q < k$, denote the smallest index such that $e_q \sim f_q$ and $e_q \succ f_{q+1}$ holds. Note that $e_{q+1} \succsim f_{q+1}$ must hold because of $B_1 \precsim_{lex} M$.
  Now we distinguish two cases:

  - $j \leq q$. Clearly we have $\sum_{f \in \tilde{B}_1 \setminus \{f_j\}} S(e_1, f) \geq l - 1 - (q-1) = l - q$ due to $e_1 \succ f_{q+1}$ and (9).
    Recall that due to the choice of q and k we have $e_j \sim f_j$, $e_{j+1} \sim f_{j+1}$, ..., $e_q \sim f_q$. Thus $e_j \sim e_{j+1} \sim ... \sim e_q$ holds (because $e_t \succ e_{t+1}$ for some $j \leq t \leq q-1$ contradicts to the choice of q since then $e_t \succ f_{t+1}$ holds as well). But this implies $f_j \sim e_q$ and $f_j \precsim e$ holds for all $e \in \{e_2, e_3, ..., e_q\}$. Hence $\sum_{e \in M_1 \setminus \{e_1\}} S(f_j, e) \leq l - 1 - (q-1) = l - q$ is satisfied as a consequence of (8). Herewith (10) holds.

– $j > q$. Because of $e_{q+1} \succsim f_{q+1}$ and (7) we get $f_j \precsim e_{q+1}$. Thus $f_j \precsim e$ holds for all $e \in \{e_2, e_3, ..., e_{q+1}\}$ which implies $\sum_{e \in M_1 \setminus \{e_1\}} S(f_j, e) \leq l-1-q$. Recall that we have $e_q \succ f_{q+1}$ because of the choice of q and thus $e_1 \succ f$ holds for all f_x with $q+1 \leq x \leq l$. This observation and (9) imply $\sum_{f \in \tilde{B}_1 \setminus \{f_j\}} S(e_1, f) \geq l-1-q$ and so (10) is satisfied.

Thus $A \succ_S M$ holds. But we have $|E_A| = |E_{B_1}| + 1$, which is a contradiction to the choice of B_1.
"$\Leftarrow$": Analogous to the proof of the corresponding direction of Theorem 4.2. □

Summarizing our results we state the following Theorem.

Theorem 4.4 *Let $M \in \tau$. Then the following statements are equivalent:*

1. *M is a max-spanning tree*
2. *$\nexists B \in \tau : B \succ_{lex} M$*
3. *$\nexists B \in \tau : B \succ_S M$*
4. *$\nexists B \in \tau : B \succ_{mxn} M$*
5. *$\nexists B \in \tau : B \succ_{ps} M$*

Proof. "1.$\Leftrightarrow$2." and "1.$\Leftrightarrow$3." were stated in Theorem 4.2 and Theorem 4.3.
The proofs of "4.$\Rightarrow$1." and "5.$\Rightarrow$1." are analogous to the one for "2.$\Rightarrow$1." (see proof of Theorem 4.2). There the assumption that M is not a max-spanning tree is led to a contradiction by creating a tree T that satisfies $T \succ_{lex} M$. But $T \succ_{mxn} M$ and $T \succ_{ps} M$ hold as well and thus an analogous contradiction can be created in either case.

"2.$\Rightarrow$4.": This can be shown analogous to direction "$\Rightarrow$" in Theorem 4.2, because due to that theorem a tree M that satisfies condition 2. is a max-spanning tree. However, we now create cycle K by adding f_r to M instead of f_{j-1}, which finally yields a tree $A_1 = B \setminus \{f_r\} \cup \{e\}$ for some $e \in M \setminus B$ with $e \succsim f_r$. Recall that $M \succeq_{lex} B$ holds, and thus $M \succ_{lex} B$ holds because $M \sim_{lex} B$ contradicts to our assumption $B \succ_{mxn} M$. $M \succ_{lex}$ implies $e_1 \succsim f_1$, and hence $e_1 \sim f_1$ must hold as $e_1 \succ f_1$ contradicts to $B \succ_{mxn} M$. $B \succ_{mxn} M$ and $f_1 \sim e_1$ imply $f_r \succ e_r$. Hence we have $e \in \{e_1, e_2, ..., e_{r-1}\}$ and $f_{r-1} \succsim f_r \succ e_r$. Thus, if $e \neq e_1$, we have $A_1 \succ_{mxn} M$ because of $f_1 \sim e_1$ and $f_{r-1} \succ e_r$. If $e = e_1$ we have $f_1 \succsim e_2$ and $f_{r-1} \succ e_r$ and so in either case we have $A_1 \succ_{mxn} M$. Repeating this procedure $r-2$ times yields a tree $A_{r-1} \succ_{mxn} M$ with $A_{r-1} \setminus M = \{f_1\}$ and $M \setminus A_{r-1} = \{e_r\}$. Thus $A_{r-1} \succ_{lex} M$ holds which is a contradiction.

"2.$\Rightarrow$5.": Analogous to "2.$\Rightarrow$4.", but now K is created by adding f_1 instead of f_{j-1}. Thus we get a tree $A_1 = B \cup \{e\} \setminus \{f_1\}$ with $e \succsim f_1$. Note that this implies $b(e) \geq b(f_1)$. Recall that $b(\tilde{B}) > b(\tilde{M})$ holds, where $\tilde{B} := B \setminus M$ and $\tilde{M} := M \setminus B$. Hence with $\tilde{A}_1 := A_1 \setminus M$ and $M_1 := M \setminus A_1$ we have $b(\tilde{A}_1) = b(\tilde{B}) - b(f_1) > b(\tilde{M}) - b(e) \geq b(M_1)$. By repeating this procedure $r-2$ times we get a tree $A_{r-1} \succ_{ps} M$ with $A_{r-1} \setminus M = \{f_r\}$ and $M \setminus A_{r-1} = \{\tilde{e}\}$ for some $\tilde{e} \in M$. This implies $A_{r-1} \succ_{lex} M$ which is a contradiction. □

Remark 1. Note that equivalence "1.$\Leftrightarrow$5." implies that the size of the numbers assigned to the edges is not crucial for the determination of a best tree with respect to $\succsim_{ps}$. I.e. for every assignment of numbers to edges according to Definition 4.3 the set of best trees w.r.t. $\succsim_{ps}$ is the same.
In order to determine the group ranking on E often positional scoring methods [6] are used. As a consequence of the above observation, every positional scoring method that yields

the same ranking $\succsim$ on E yields the same set of best trees w.r.t. $\succsim_{ps}$, irrespective of the numerical values assigned to the edges.

Remark 2. Due to the fact that a max-spanning tree can be determined efficiently by applying a greedy algorithm, for all orders regarded in Theorem 4.4 a socially best tree can be computed in polynomial time.

5 Conclusion

In this paper we have presented different ways to achieve orderings on the set of spanning trees from a group ranking on the edge-set of a graph. Assuming that the given group ranking is a weak order (which does not necessarily allocate a numerical value to each edge), we have shown that the sets of socially best trees according to the concepts discussed in this paper coincide.

In the related work of Perny and Spanjaard [9] a quite general framework for ranking spanning trees on the basis of preference relations is presented. In [9] a main focus is laid on establishing sufficient conditions for an order on the power set of the edge-set under which a greedy algorithm is able to determine the set of best trees with respect to the corresponding order. The orders $\succsim_{lex}$, $\succsim_S$, $\succsim_{mxn}$ and $\succsim_{ps}$ however[2] do not belong to the class of orders for which these conditions hold, even if the group ranking is assumed to be a partial order on the set of edges. As a consequence of our paper, a socially best tree which respect to each of these orders can be computed efficiently by simply determining a max-spanning tree using a greedy algorithm.

[2] if their definitions adequately are extended to the power set of the edge-set

References

[1] R.K. Ahuja, T.L. Magnanti, and J.B. Orlin. *Network flows: theory, algorithms, and applications*. Prentice Hall, 1993.

[2] S. Barbera, W. Bossert, and P.K. Pattanaik. Ranking sets of objects. In *Handbook of Utility Theory*, volume 2, pages 893–977. 2004.

[3] J.J. Bartholdi, C.A. Tovey, and M.A. Trick. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare*, 6:157–165, 1989.

[4] M.W. Bern and D. Eppstein. Worst-case bounds for subadditive geometric graphs. In *Symposium on Computational Geometry*, pages 183–188, 1993.

[5] D. Black. *The Theory of Committees and Elections*. Cambridge University Press, Cambridge, 1958.

[6] S.J. Brams and P.C. Fishburn. Voting procedures. In K.J. Arrow, A.K. Sen, and K. Suzumura, editors, *Handbook of Social Choice and Welfare*, volume 1, pages 173–236. Elsevier, 2002.

[7] A.H. Copeland. *A 'Reasonable' Social Welfare Function*. Notes from a Seminar on Applications of Mathematics to the Social Sciences, University of Michigan, 1951.

[8] H. Nurmi. *Voting Procedures under Uncertainty*. Springer Verlag, Berlin, 2007.

[9] P. Perny and O. Spanjaard. A preference-based approach to spanning trees and shortest paths problems. *European Journal of Operational Research*, 127(3):584–601, May 2005.

[10] D.G. Saari. *Basic Geometry of Voting*. Springer Verlag, Berlin, 1995.

[11] P. Slater. Inconsistencies in a schedule of paired comparisons. *Biometrica*, 48:303–312, 1961.

Andreas Darmann, Christian Klamler
Institute of Public Economics
University of Graz
8010 Graz, Austria
Email: {andreas.darmann,christian.klamler}@uni-graz.at

Ulrich Pferschy
Institute of Statistics and Operations Research
University of Graz
8010 Graz, Austria
Email: pferschy@uni-graz.at

Majority voting on restricted domains: a summary

Franz Dietrich and Christian List

Abstract. In judgment aggregation, unlike preference aggregation, not much is known about domain restrictions that guarantee consistent majority outcomes. We introduce several conditions on individual judgments sufficient for consistent majority judgments. Some are based on global orders of propositions or individuals, others on local orders, still others not on orders at all. Some generalize classic social-choice-theoretic domain conditions, others have no counterpart. Our most general condition generalizes Sen's triplewise value-restriction, itself the most general classic condition. Taken together, our results suggest that majority inconsistencies can be avoided in practice, provided that disagreements are appropriately structured. This rehabilitates majority voting as a potential way to reach collective judgments.

1 Introduction

In the theory of preference aggregation, it is well known that majority voting on pairs of alternatives may generate inconsistent (i.e., cyclical) majority preferences even when all individuals' preferences are consistent (i.e., acyclical). The most famous example is Condorcet's paradox. Here one individual prefers x to y to z, a second y to z to x, and a third z to x to y, and thus there are majorities for x against y, for y against z, and for z against x, a 'cycle'. But it is equally well known that if individual preferences fall into a suitably restricted domain, majority cycles can be avoided (for an excellent overview, see Gaertner [16]). The most famous domain restriction with this effect is Black's single-peakedness [1]. A profile of individual preferences is single-peaked if the alternatives can be ordered from 'left' to 'right' such that each individual has a most preferred alternative with decreasing preference for other alternatives as we move away from it in either direction. Inada [18] showed that another condition called single-cavedness and interpretable as the mirror image of single-peakedness also suffices for avoiding majority cycles: a profile is single-caved if, for some 'left'-'right' order of the alternatives, each individual has a least preferred alternative with increasing preference for other alternatives as we move away from it in either direction. Sen [37] introduced a very general domain restriction, called triplewise value-restriction, that garantees acyclical majority preferences and is implied by Black's, Inada's and other conditions; it therefore unifies several domain-restriction conditions, yet has a technical flavour without straightforward interpretation.

The wealth of domain-restriction conditions for avoiding majority cycles was supplemented by another family of conditions based not on 'left'-'right' orders of the alternatives, but on 'left'-'right' orders of the individuals. Important conditions in this family are Grandmont's intermediateness [17] and Rothstein's order restriction ([34], [35]) with its special case of single-crossingness (e.g., Saporiti and Tohmé [36]). To illustrate, a profile of individual preferences is order-restricted if the individuals – rather than the alternatives – can be ordered from 'left' to 'right' such that, for each pair of alternatives x and y, the individuals preferring x to y are either all to the left, or all the right, of those preferring y to x.

Empirically, domain restrictions are important as many political and economic contexts induce a natural structure in preferences. For example, domain restrictions based on a 'left'-'right' order – whether of the alternatives or of the individuals – can capture situations in which preferences are structured by one normative or cognitive dimension, such as from socialist to libertarian, from urban to rural, or from secular to religious.

In the theory of judgment aggregation, by contrast, domain restrictions have received much less attention (the only exception is the work on unidimensional alignment, e.g., List [22]). This is an important gap in the literature since, here too, majority voting with unrestricted but consistent individual inputs may generate inconsistent collective outputs, while on a suitably restricted domain such inconsistencies can be avoided. As illustrated by the much-discussed discursive paradox (e.g., Pettit [31]), if one individual judges that a, $a \rightarrow b$ and b, a second that a, but not $a \rightarrow b$ and not b, and a third that $a \rightarrow b$, but not a and not b, there are majorities for a, for $a \rightarrow b$ and yet for not b, an inconsistency. But if no individual rejects $a \rightarrow b$, for example, this problem can never arise.

Surprisingly, however, despite the abundance of impossibility results generalizing the discursive paradox as reviewed below, very little is known about the domains of individual judgments on which discursive paradoxes can occur (as opposed to agendas of propositions susceptible to such problems, which have been extensively characterized in the literature). If we can find compelling domain restrictions to ensure majority consistency, this allows us to refine and possibly amelioriate the lessons of the discursive paradox. Going beyond the standard impossibility results, which all assume an unrestricted domain, we can then ask: in what political and economic contexts do the identified domain restrictions hold, so that majority voting becomes safe, and in what contexts are they violated, so that majority voting becomes problematic?

This paper introduces several conditions on profiles of individual judgments that guarantee consistent majority judgments. As explained in a moment, these can be distinguished in at least two respects: first, in terms of whether they are based on orders of propositions, on orders of individuals, or not on orders at all; and second, if they are based on orders, in terms of whether these are 'global' or 'local'. We also discuss parallels and disanalogies with domain-restriction conditions on preferences.

Let us briefly comment on the two distinctions underlying our discussion. First, our conditions based on orders of the individuals are analogous to, and in fact generalize, some of the conditions on preferences reviewed above, particularly intermediateness and order restriction. By contrast, those conditions based on orders of the propositions are not obviously analogous to any conditions on preferences. While an order of individuals can be interpreted similarly in judgment and preference aggregation – namely in terms of the individuals' positions on a normative or cognitive dimension – an order of propositions in judgment aggregation is conceptually distinct from an order of alternatives in preference aggregation. Propositions, unlike alternatives, are not mutually exclusive. It is therefore surprising that sufficient conditions for consistent majority judgments can be given even based on orders of propositions. We also introduce a very general domain-restriction condition not based on orders at all: it generalizes Sen's condition of triplewise value-restriction. In concluding the paper, we characterize the maximal domain on which majority voting yields consistent collective judgments.

Secondly, our domain-restriction conditions based on orders admit global and local variants. In the global case, the individuals' judgments on all propositions on the agenda are constrained by the same 'left'-'right' order of propositions or individuals, whereas in the local case, that order may differ across subsets of the agenda. To give an illustration from the more familiar context of preference aggregation, single-peakedness and single-cavedness are global conditions, whereas the restriction of these conditions to triples of alternatives yields local ones. But while in preference aggregation local conditions result from the restriction of global conditions to triples of alternatives, the picture is more general in judgment aggregation. Here different 'left'-'right' orders may apply to different subagendas, which correspond to different semantic fields. We give precise criteria for selecting appropriate subagendas. An individual can be left-wing on a 'social' subagenda and right-wing on an 'economic' one, for example.

As already noted, some of our conditions generalize existing conditions in preference aggregation, notably Grandmont's intermediateness, Rothstein's order restriction and Sen's triplewise value-restriction, and reduce to them when the agenda of propositions under consideration contains binary ranking propositions suitable for representing preferences (such as xPy, yPz, xPz etc.).[1]

We state our results for the general case in which individual and collective judgments are only required to be consistent; they need not be complete (i.e., they need not be opinionated on every proposition-negation pair). But whenever this is relevant, we also consider the important special case of full rationality, i.e., the conjunction of consistency and completeness.

A few remarks about the literature on judgment aggregation are due. The recent field of judgment aggregation emerged from the areas of law and political philosophy (e.g., Kornhauser and Sager [20] and Pettit [31]) and was formalized social-choice-theoretically by List and Pettit [24]. The literature contains several impossibility results generalizing the observation that on an unrestricted domain majority judgments can be logically inconsistent (e.g., List and Pettit [24] and [25], Pauly and van Hees [30], Dietrich [2], Gärdenfors [15], Nehring and Puppe [29], van Hees [38], Mongin [26], Dietrich and List [7], and Dokow and Holzman [13]). Other impossibility results follow from Nehring and Puppe's [27] strategy-proofness results on property spaces. Earlier precursors include works on abstract aggregation (Wilson [39], Rubinstein and Fishburn [33]). A liberal-paradox-type impossibility was derived in Dietrich and List [12]. Giving up propositionwise aggregation, possibility results were obtained, for example, by using sequential rules (List [23]) and fusion operators (Pigozzi [32]). Voter manipulation in the judgment-aggregation model was analysed in Dietrich and List [8]. But so far the only domain-restriction condition known to guarantee consistent majority judgments is List's unidimensional alignment ([21], [22]), a global domain condition based on orders of individuals. Here we use Dietrich's generalized model [3], which allows propositions to be expressed in rich logical languages.

This paper summarizes results from our working paper [6], in which we also give proofs; these are omitted here for brevity. We are grateful to the ComSoc referees for comments and suggestions.

2 The model

We consider a group of individuals $N = \{1, 2, \ldots, n\}$ ($n \geq 2$) making judgments on some propositions represented in logic (Dietrich [3], generalizing List and Pettit [24], [25]).

Logic. A *logic* is given by a *language* and a notion of *consistency*. The *language* is a non-empty set $\mathbf{L}$ of sentences (called *propositions*) closed under negation (i.e., $p \in \mathbf{L}$ implies $\neg p \in \mathbf{L}$, where $\neg$ is the negation symbol). For example, in standard propositional logic, $\mathbf{L}$ contains propositions such as a, b, $a \wedge b$, $a \vee b$, $\neg(a \rightarrow b)$ (where $\wedge$, $\vee$, $\rightarrow$ denote 'and', 'or', 'if-then', respectively). In other logics, the language may involve additional connectives, such as modal operators ('it is necessary/possible that'), deontic operators ('it is obligatory/permissible that'), subjunctive conditionals ('if p were the case, then q would be the case'), or quantifiers ('for all/some'). The notion of *consistency* captures the logical connections between propositions by stipulating that some sets of propositions $S \subseteq \mathbf{L}$ are *consistent* (and the others *inconsistent*), subject to some regularity axioms.[2] A proposition $p \in \mathbf{L}$ is a *contradiction* if $\{p\}$ is inconsistent and a *tautology* if $\{\neg p\}$ is inconsistent. For example,

[1] The fact that these three existing conditions are already very general representatives of their respective families underlines the generality of our new conditions here.

[2] Self-entailment: Any pair $\{p, \neg p\} \subseteq \mathbf{L}$ is inconsistent. Monotonicity: Subsets of consistent sets $S \subseteq \mathbf{L}$ are consistent. Completability: $\emptyset$ is consistent, and each consistent set $S \subseteq \mathbf{L}$ has a consistent superset $T \subseteq \mathbf{L}$ containing a member of each pair $p, \neg p \in \mathbf{L}$. See Dietrich [3].

in standard logics, $\{a, a \to b, b\}$ and $\{a \wedge b\}$ are consistent and $\{a, \neg a\}$ and $\{a, a \to b, \neg b\}$ inconsistent; $a \wedge \neg a$ is a contradiction and $a \vee \neg a$ a tautology.

Agenda. The *agenda* is the set of propositions on which judgments are to be made. It is a non-empty set $X \subseteq \mathbf{L}$ expressible as $X = \{p, \neg p : p \in X_+\}$ for some set X_+ of unnegated propositions (this avoids double-negations in X). In our introductory example, the agenda is $X = \{a, \neg a, a \to b, \neg(a \to b), b, \neg b\}$. For convenience, we assume that X is finite.[3] As a notational convention, we cancel double-negations in front of propositions in X.[4] Further, for any $Y \subseteq X$, we write $Y^{\pm} = \{p, \neg p : p \in Y\}$ to denote the (single-)negation closure of Y.

Judgment sets. An individual's *judgment set* is the set $A \subseteq X$ of propositions in the agenda that he or she accepts (e.g., 'believes'). A *profile* is an n-tuple $(A_1, \ldots, A_n)$ of judgment sets across individuals. A judgment set is *consistent* if it is consistent in **L;** it is *complete* if it contains at least one member of each proposition-negation pair $p, \neg p \in X$; it is *opinionated* if it contains precisely one such member. Our results mostly do not require completeness, in line with several works on the aggregation of incomplete judgments (Gärdenfors [15]; Dietrich and List [9], [10], [11]; Dokow and Holzman [14]; List and Pettit [24]). This strengthens our possibility results as the identified possibilities hold on larger domains of profiles. But we also consider the complete case.

Aggregation functions. A *domain* is a set D of profiles, interpreted as admissible inputs to the aggregation. An *aggregation function* is a function F that maps each profile $(A_1, \ldots, A_n)$ in a given domain D to a collective judgment set $F(A_1, \ldots, A_n) = A \subseteq X$. While the literature focuses on the *universal domain* (which consists of all profiles of consistent and complete judgment sets), we here focus mainly on domains that are less restrictive in that they allow for incomplete judgments, but more restrictive in that we impose some structural conditions. We call an aggregation function *consistent* or *complete*, respectively, if it generates a consistent or complete judgment set for each profile in its domain. The *majority outcome on a profile* $(A_1, ..., A_n)$ is the judgment set $\{p \in X :$ there are more individuals $i \in N$ with $p \in A_i$ than with $p \notin A_i\}$. The aggregation function that generates the majority outcome on each profile in its domain D is called *majority voting on* D.[5]

Preference aggregation as a special case. To relate our results to existing results on preference aggregation, we must explain how preference aggregation can be represented in our model.[6] Since preference relations are binary relations on some set, they allow a logical representation. Take a simple predicate logic **L** with a set of two or more constants $K = \{x, y, ...\}$ representing alternatives and a two-place predicate P representing (strict) preference. For any $x, y \in K$, xPy means 'x is preferable to y'. Define any set $S \subseteq \mathbf{L}$ to be *consistent* if $S \cup Z$ is consistent in the standard sense of predicate logic, where Z is the set of rationality conditions on strict preferences.[7] Now the *preference agenda* is $X_K = \{xPy \in \mathbf{L} : x, y \in K\}^{\pm}$. Preference relations and opinionated judgment sets stand in a bijective correspondence:

[3] For infinite X, our results hold either as stated or under compactness of the logic.

[4] More precisely, if $p \in X$ is already of the form $p = \neg q$, we write $\neg p$ to mean q rather than $\neg\neg q$. This ensures that, whenever $p \in X$, then $\neg p \in X$.

[5] Other widely discussed aggregation functions include dictatorships, supermajority functions, and premise-based or conclusion-based functions.

[6] For details of the construction, see Dietrich and List [7], extending List and Pettit [25].

[7] Z consists of $(\forall v_1)(\forall v_2)(v_1 P v_2 \to \neg v_2 P v_1)$ (asymmetry), $(\forall v_1)(\forall v_2)(\forall v_3)((v_1 P v_2 \wedge v_2 P v_3) \to v_1 P v_3)$ (transitivity), $(\forall v_1)(\forall v_2)(\neg\, v_1 = v_2 \to (v_1 P v_2 \vee v_2 P v_1))$ (connectedness) and, for each pair of distinct constants $x, y \in K$, $\neg\, x = y$ (exclusiveness of alternatives).

- to any preference relation (arbitrary binary relation) $\succ$ on K corresponds the opinionated judgment set $A_\succ \subseteq X_K$ with $A_\succ = \{xPy : x, y \in K \& x \succ y\} \cup \{\neg xPy : x, y \in K \& x \not\succ y\}$;
- conversely, to any opinionated judgment set $A \subseteq X_K$ corresponds the preference relation $\succ_A$ on K with $x \succ_A y \Leftrightarrow xPy \in A \; \forall x, y \in K$.

A preference relation $\succ$ is fully rational (i.e., asymmetric, transitive and connected) if and only if $A_\succ$ is consistent, because we have built the rationality conditions on preferences into the logic. Under this construction, a judgment aggregation function (for opinionated judgment sets) represents a preference aggregation function, and majority voting as defined above corresponds to pairwise majority voting in the standard Condorcetian sense.

3 Conditions for majority consistency based on global orders

On which domains of profiles is majority voting consistent? We already know from the discursive paradox that without any domain restriction it is not (unless the agenda is trivial).[8] However, we now show that there exist many compelling domains on which majority voting is consistent.

3.1 Conditions based on orders of propositions

We begin with two conditions based on 'global' orders of the propositions. An *order of the propositions (in X)* is a linear order $\leq$ on X.[9]

Single-plateauedness. A judgment set A is *single-plateaued relative to* $\leq$ if $A = \{p \in X : p_{\text{left}} \leq p \leq p_{\text{right}}\}$ for some $p_{\text{left}}, p_{\text{right}} \in X$, and a profile is $(A_1, ..., A_n)$ is *single-plateaued relative to* $\leq$ if every A_i is *single-plateaued* relative to $\leq$.

Single-canyonedness. A judgment set A is *single-canyoned relative to* $\leq$ if $A = X \backslash \{p \in X : p_{\text{left}} \leq p \leq p_{\text{right}}\}$ for some $p_{\text{left}}, p_{\text{right}} \in X$, and a profile is $(A_1, ..., A_n)$ is *single-canyoned relative to* $\leq$ if every A_i is *single-canyoned* relative to $\leq$.[10]

An order $\leq$ that renders a profile single-plateaued or single-canyoned is called a *structuring order*; it need not be unique. If a profile is single-plateaued or single-canyoned relative to some $\leq$, we also call it *single-plateaued* or *single-canyoned* simpliciter. Both conditions are illustrated in Figure 1.

The order $\leq$ may represent a normative or cognitive dimension on which propositions are located. If the agenda contains scientific propositions about global warming, for example, individuals may hold single-plateaued judgment sets relative to an order of the propositions from 'most pessimistic' to 'most optimistic', and the location of each individual's plateau may reflect his or her scientific position. If the agenda contains propositions about the effects of various tax or budget policies, the propositions may be ordered from 'socialist' to 'libertarian'. If the agenda contains propositions concerning biological issues, the order may

[8] Majority inconsistencies can arise whenever the agenda has a minimal inconsistent subset of three or more propositions. For a proof of this fact under consistency alone, see Dietrich and List [9]; under full rationality, see Nehring and Puppe [28].

[9] Thus $\leq$ is reflexive ($x \leq x \; \forall x$), transitive ($[x \leq y$ and $y \leq z] \Rightarrow x \leq z \; \forall x, y, z$), connected ($x \neq y \Rightarrow [x \leq y$ or $y \leq x] \; \forall x, y$) and antisymmetric ($[x \leq y$ and $y \leq x] \Rightarrow x = y \; \forall x, y$).

[10] In the definitions of single-plateauedness and single-canyonedness, we do not require $p_{\text{left}} \leq p_{\text{right}}$, i.e., $\{p : p_{\text{left}} \leq p \leq p_{\text{right}}\}$ may be empty.

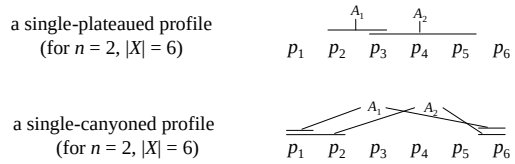


Figure 1: Single-plateauedness and single-canyonedness

range from 'closest to theory X' (e.g., evolutionary theory) to 'closest to theory Y' (e.g., creationism).

We first observe that every single-canyoned profile is single-plateaued.

Proposition 1 *Every single-canyoned profile* $(A_1, ..., A_n)$ *of consistent judgment sets is single-plateaued.*

As anticipated, majority voting preserves consistency on single-plateaued profiles. On single-canyoned profiles, it does even more: it also preserves single-canyonedness.

Proposition 2 *For any profile* $(A_1, ..., A_n)$ *of consistent judgment sets,*
(a) if $(A_1, ..., A_n)$ *is single-plateaued, the majority outcome is consistent;*
(b) if $(A_1, ..., A_n)$ *is single-canyoned, the majority outcome is consistent and single-canyoned (relative to the same structuring order).*

3.2 Conditions based on orders of individuals

Let us now turn to two conditions based on 'global' orders of the individuals. An *order of the individuals (in N)* is linear order Ω on N. For any sets of individuals $N_1, N_2 \subseteq N$, we write $N_1 \Omega N_2$ if $i \Omega j$ for all $i \in N_1$ and $j \in N_2$.

Unidimensional orderedness.[11] A profile $(A_1, ..., A_n)$ is *unidimensionally ordered relative to* Ω if, for all $p \in X$, $\{i \in N : p \in A_i\} = \{i \in N : i_{\text{left}} \Omega i \Omega i_{\text{right}}\}$ for some $i_{\text{left}}, i_{\text{right}} \in N$.

Unidimensional alignment. (List [22]) A profile $(A_1, ..., A_n)$ is *unidimensionally aligned relative to* Ω if, for all $p \in X$, $\{i \in N : p \in A_i\} \Omega \{i \in N : p \notin A_i\}$ or $\{i \in N : p \notin A_i\} \Omega \{i \in N : p \in A_i\}$.

In analogy to the earlier definition, an order Ω that renders a profile unidimensionally ordered or unidimensionally aligned is called a *structuring order*; again, it need not be unique. If a profile is unidimensionally ordered or unidimensionally aligned relative to some Ω, we also call it *unidimensionally ordered* or *unidimensionally aligned* simpliciter. Both conditions are illustrated in Figure 2.

Unidimensional alignment is a special case of unidimensional orderedness: it is the case in which, for every $p \in X$, at least one of $i_{\text{left}}, i_{\text{right}}$ is 'extreme', i.e., the left-most or right-most individual in the structuring order Ω.

Proposition 3 *Every unidimensionally aligned profile* $(A_1, ..., A_n)$ *is unidimensionally ordered.*

How can we interpret the two conditions? A profile is unidimensionally ordered if the individuals can be ordered from 'left' to 'right' such that, for each proposition, the individuals accepting it are all adjacent to each other; a profile is unidimensionally aligned if, in

[11] In this definition, we do not require $i_{\text{left}} \Omega i_{\text{right}}$, i.e., $\{i : i_{\text{left}} \Omega i \Omega i_{\text{right}}\}$ may be empty.

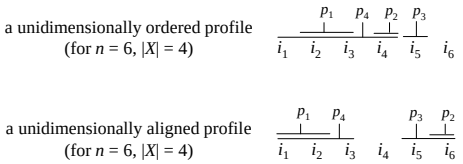


Figure 2: Unidimensional orderedness and unidimensional alignment

addition, the individuals accepting each proposition are either all to the left or all to right of those rejecting it. The order of the individuals can be interpreted as reflecting their location on some underlying normative or cognitive dimension. The idea underlying unidimensional orderedness is that each proposition, like each individual, is located somewhere on the dimension and is accepted by those individuals whose location is 'close' to it, hence by some interval of individuals 'around' it. In a decision problem about climate policies, for example, the proposition 'taxation on emissions should be moderately increased' might have a central location and might therefore be accepted by a 'central' interval of individuals. In the case of unidimensional alignment, the extreme positions on the given dimension correspond to either clear acceptance or clear rejection of each proposition, and, for each proposition, there is a threshold between these extremes (which may vary across propositions) that divides the 'acceptance-region' from the 'rejection-region'.[12]

On unidimensionally ordered profiles, majority voting preserves consistency, and we can say something about the nature of its outcome: it is a subset of the middle individual's judgment set (or, for even n, a subset of the intersection of the two middle individuals' judgment sets). If the profile is unidimensionally aligned, the majority outcome is not just included in that set but coincides with it.

Proposition 4 *For any profile $(A_1, ..., A_n)$ of consistent judgment sets,*

(a) if $(A_1, ..., A_n)$ is unidimensionally ordered, the majority outcome A is consistent and

$$A \subseteq \begin{cases} A_m & \text{if } n \text{ is odd,} \\ A_{m_1} \cap A_{m_2} & \text{if } n \text{ is even,} \end{cases}$$

where m is the middle individual (if n is odd) and m_1, m_2 the middle pair of individuals (if n is even) in any structuring order Ω;

(b) (List [22]) if $(A_1, ..., A_n)$ is unidimensionally aligned, the majority outcome is as stated in part (a) with $\subseteq$ replaced by $=$.

3.3 The logical relationships between the four conditions

We have already seen that single-canyonedness implies single-plateauedness, and that unidimensional alignment implies unidimensional orderedness. A natural question is how the first two conditions, which are based on orders of the propositions, are related to the second two, which are based on orders of the individuals. The following result answers this question.

Proposition 5 *(a) Restricted to profiles of consistent judgment sets,*

- *unidimensional alignment implies any of the other three conditions;*
- *single-canyonedness implies single-plateauedness;*
- *there are no other pairwise implications between the four conditions.*

(b) Restricted to profiles of consistent and complete (or just of opinionated) judgment sets, the four conditions are equivalent.

[12] In List [21], unidimensional alignment is interpreted in terms of 'meta-agreement'.

3.4 Applications to preference aggregation: order restriction and intermediateness

As we explain precisely in Dietrich and List [6], the conditions based on orders of the individuals reduce to classic conditions if applied to the preference agenda:

Remark 6 (a) *A preference profile* $(\succ_1, ..., \succ_n)$ *is order restricted (relative to some* Ω*) if and only if the corresponding judgment profile* $(A_{\succ_1}, ..., A_{\succ_n})$ *is unidimensionally aligned (relative to the same* Ω*).*

(b) *An opinionated preference profile* $(\succ_1, ..., \succ_n)$ *is intermediate (relative to some* Ω*) if and only if the corresponding judgment profile* $(A_{\succ_1}, ..., A_{\succ_n})$ *is unidimensionally ordered (relative to the same* Ω*), where opinionation means that, for each* $i \in N$ *and all distinct* $x, y \in K$*, precisely one of* $x \succ_i y$ *or* $y \succ_i x$ *holds.*

4 Conditions for majority consistency based on local orders

For many agendas, the four domain-restriction conditions discussed so far are stronger than necessary for achieving majority consistency. As developed in detail in our working paper [6], it suffices to apply our conditions to the judgments on various subagendas of X, thereby allowing the relevant structuring order of individuals or propositions to vary across different subagendas. This move parallels the move in preference aggregation from single-peakedness to single-peakedness restricted to triples of alternatives.

Formally, a *subagenda (of* X*)* is a subset $Y \subseteq X$ that is itself an agenda (i.e., non-empty and closed under single negation). For each of our four global domain-restriction conditions, we say that a profile $(A_1, ..., A_n)$ satisfies the given condition *on a subagenda* $Y \subseteq X$ if the restricted profile $(A_1 \cap Y, ..., A_n \cap Y)$, viewed as a profile of judgment sets on the agenda Y, satisfies it. The relevant structuring order is then called a *structuring order on* Y and denoted $\leq_Y$ (if it is an order of propositions) or Ω_Y (if it is an order of individuals). Whenever one of the conditions is satisfied globally, then it is also satisfied on every $Y \subseteq X$. But we now define a local counterpart of each global condition. Let $\mathcal{Y}$ be some set of subagendas.

Local single-plateauedness / single-canyonedness / unidimensional orderedness / unidimensional alignment. A profile $(A_1, ..., A_n)$ satisfies *the local counterpart of each global condition* (with respect to a given set of subagendas $\mathcal{Y}$) if it satisfies the global condition on every $Y \in \mathcal{Y}$.

Provided the set of subagendas $\mathcal{Y}$ is suitably chosen, these local conditions are sufficient to ensure consistent majority outcomes (if individuals hold consistent judgments). In our working paper [6], we discuss two choices of subagendas; according to the first specification,

$$\mathcal{Y} = \{Y^{\pm} : Y \text{ is a minimal inconsistent subset of } X\}.$$

The second specification uses so-called irreducible sets and generalizes the classic local conditions of *intermediateness on triples* and *order restriction on triples* in preference aggregation.

5 Conditions for majority consistency not based on orders

Although our domain-restriction conditions based on local orders are already much less restrictive than those based on global orders, it is possible to weaken them further. Just as the various conditions based on orders in preference aggregation – single-peakedness, single-cavedness etc. – can be generalized to a weaker, but less easily interpretable, condition – namely Sen's triplewise value-restriction [37] – so in judgment aggregation the conditions based on orders can be weakened to a more abstract condition, to be called *value-restriction*. When applied to the preference agenda, this condition becomes non-trivially equivalent to Sen's condition. But despite generalizing Sen's condition, our condition is simpler to state; we thus also hope to offer a new perspective on Sen's condition.

5.1 Value-restriction

For any inconsistent set $Y \subseteq X$, we call another inconsistent set $Z \subseteq X$ a *reduction* of Y if

$$|Z| < |Y| \text{ and each } p \in Z \backslash Y \text{ is entailed by some } V \subseteq Y \text{ with } |Y \backslash V| > 1,$$

and we call Y *irreducible* if it has no reduction. For instance, the inconsistent set $\{a, a \rightarrow b, b \rightarrow c, \neg c\}$ (where a, b, c are distinct atomic propositions) is reducible to $Z = \{b, b \rightarrow c, \neg c\}$, since b is entailed by $\{a, a \rightarrow b\}$, whereas Z is irreducible.

We state two variants of our condition, one based on minimal inconsistent sets, the other based on irreducible sets.

Value-restriction. A profile $(A_1, ..., A_n)$ is *value-restricted* if every (non-singleton[13]) minimal inconsistent set $Y \subseteq X$ has a two-element subset $Z \subseteq Y$ that is not a subset of any A_i.

Weak value-restriction. A profile $(A_1, ..., A_n)$ is *weakly value-restricted* if every (non-singleton) irreducible set $Y \subseteq X$ has a two-element subset $Z \subseteq Y$ that is not a subset of any A_i.

Informally, value-restriction reflects a particular kind of agreement: for every minimal inconsistent (or irreducible in the weak case) subset of the agenda, there exists a particular conjunction of two propositions in this subset that no individual endorses. Like our previous domain-restriction conditions, the two new conditions are each sufficient for consistent majority outcomes (the weaker condition in the important special case of individual completeness).

Proposition 7 *For any profile $(A_1, ..., A_n)$ of consistent judgment sets,*
(a) if $(A_1, ..., A_n)$ is value-restricted, the majority outcome is consistent;
(b) if $(A_1, ..., A_n)$ is weakly value-restricted and each A_i is complete, the majority outcome is consistent.

How general are our two value-restriction conditions? The following proposition answers this question.

Proposition 8 *(a) Each of our four conditions based on global orders implies value-restriction.*

[13] The qualification 'non-singleton' in this definition and the next is unnecessary if X contains only contingent propositions, since this rules out singleton inconsistent sets.

(b) Each of our four conditions based on local orders, with respect to $\mathcal{Y}$ defined in terms of minimal inconsistent sets, implies value-restriction.

(c) Each of our four conditions based on local orders, with respect to $\mathcal{Y}$ defined in terms of irreducible sets, implies weak value-restriction.

5.2 Applications to preference aggregation: triplewise value-restriction

When applied to the preference agenda, our two value-restriction conditions surprisingly both collapse into Sen's ([37]) triplewise value-restriction.

Proposition 9 *For any profile $(A_1, ..., A_n)$ of consistent and complete judgment sets on the preference agenda, the following are equivalent:*

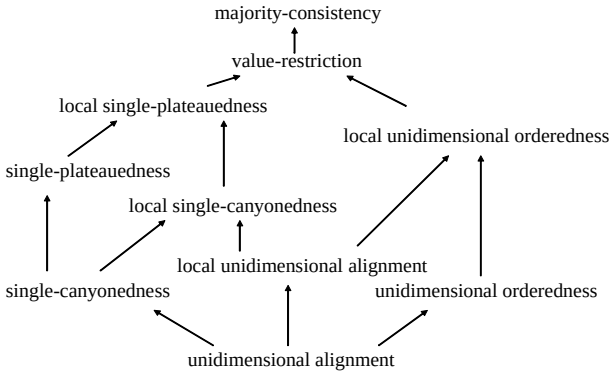
(a) $(A_1, ..., A_n)$ is value-restricted,

(b) $(A_1, ..., A_n)$ is weakly value-restricted,

(c) the associated preference profile $(\succ_{A_1}, ..., \succ_{A_n})$ is triplewise value-restricted.

6 Conclusion

The following figure summarizes the logical relationship between all the domain-restriction conditions discussed in this paper, in each case applied to profiles of consistent individual judgment sets.



References

[1] Black, D. (1948) On the Rationale of Group Decision-Making. *J Polit Economy* 56: 23-34

[2] Dietrich, F. (2006) Judgment Aggregation: (Im)Possibility Theorems. *J Econ Theory* 126(1): 286-298

[3] Dietrich, F. (2007) A generalised model of judgment aggregation. *Soc Choice Welfare* 28(4): 529-565

[4] Dietrich, F. (forthcoming) The possibility of judgment aggregation on agendas with subjunctive implications. *J Econ Theory*

[5] Dietrich, F., List, C. (2005) The impossibility of unbiased judgment aggregation. Working paper, London School of Economics

[6] Dietrich, F., List, C. (2007) Majority voting on restricted domains. Working paper, London School of Economics

[7] Dietrich, F., List, C. (2007) Arrow's theorem in judgment aggregation. *Soc Choice Welfare* 29(1): 19-33

[8] Dietrich, F., List, C. (2007) Strategy-proof judgment aggregation. *Econ Philos* 23(3)

[9] Dietrich, F., List, C. (2007) Judgment aggregation by quota rules: majority voting generalized. *J Theor Politics* 19(4): 391-424

[10] Dietrich, F., List, C. (forthcoming) Judgment aggregation without full rationality. *Soc Choice Welfare*

[11] Dietrich, F., List, C. (2007) Judgment aggregation with consistency alone. Working paper, London School of Economics

[12] Dietrich, F., List, C. (forthcoming) A liberal paradox for judgment aggregation. *Soc Choice Welfare*

[13] Dokow, E., Holzman, R. (forthcoming) Aggregation of binary evaluations. *J Econ Theory*

[14] Dokow, E., Holzman, R. (2006) Aggregation of binary evaluations with abstentions. Working paper, Technion Israel Institute of Technology

[15] Gärdenfors, P. (2006) An Arrow-like theorem for voting with logical consequences. *Econ Philos* 22(2): 181-190

[16] Gaertner, W. (2001) *Domain Conditions in Social Choice Theory.* Cambridge (Cambridge University Press)

[17] Grandmont, J.-M. (1978) Intermediate Preferences and the Majority Rule. *Econometrica* 46(2): 317-330

[18] Inada, K.-I. (1964) A Note on the Simple Majority Decision Rule. *Econometrica* 32(4): 525-531

[19] Elsholtz, C., List, C. (2005) A Simple Proof of Sen's Possibility Theorem on Majority Decisions. *Elemente der Mathematik* 60: 45-56

[20] Kornhauser, L. A., Sager, L. G. (1986) Unpacking the Court. *Yale Law Journal* 96(1): 82-117

[21] List, C. (2002) Two Concepts of Agreement. *The Good Society* 11(1): 72-79

[22] List, C. (2003) A Possibility Theorem on Aggregation over Multiple Interconnected Propositions. Mathematical Social Sciences 45(1): 1-13 (Corrigendum in *Mathematical Social Sciences* 52:109-110)

[23] List, C. (2004) A Model of Path-Dependence in Decisions over Multiple Propositions. *Amer Polit Science Rev* 98(3): 495-513

[24] List, C., Pettit, P. (2002) Aggregating Sets of Judgments: An Impossibility Result. *Econ Philos* 18: 89-110

[25] List, C., Pettit, P. (2004) Aggregating Sets of Judgments: Two Impossibility Results Compared. *Synthese* 140(1-2): 207-235

[26] Mongin, P. (forthcoming) Factoring out the impossibility of logical aggregation. *J Econ Theory*

[27] Nehring, K., Puppe, C. (2002) Strategy-Proof Social Choice on Single-Peaked Domains: Possibility, Impossibility and the Space Between. Working paper, University of California at Davies

[28] Nehring, K., Puppe, C. (2007) Abstract Arrovian Aggregation. Working paper, University of Karlsruhe

[29] Nehring, K., Puppe, C. (forthcoming) Consistent judgement aggregation: the truth-functional case. *Soc Choice Welfare*

[30] Pauly, M., van Hees, M. (2006) Logical Constraints on Judgment Aggregation. *J Philos Logic* 35: 569-585

[31] Pettit, P. (2001) Deliberative Democracy and the Discursive Dilemma. *Philosophical Issues* 11: 268-299

[32] Pigozzi, G. Belief merging and the discursive dilemma: an argument-based account to paradoxes of judgment aggregation. *Synthese* 152(2): 285-298

[33] Rubinstein, A., Fishburn, P. (1986) Algebraic Aggregation Theory. *J Econ Theory* 38: 63-77

[34] Rothstein, P. (1990) Order Restricted Preferences and Majority Rule. *Soc Choice Welfare* 7(4): 331-342

[35] Rothstein, P. (1991) Representative Voter Theorems. *Public Choice* 72(2-3): 193-212

[36] Saporiti, A., Tohmé, F. (2006) Single-crossing, strategic voting and the median choice rule. *Soc Choice Welfare* 26(2): 363-383

[37] Sen, A. K. (1966) A Possibility Theorem on Majority Decisions. *Econometrica* 34(2): 491-499

[38] van Hees, M. (2007) The limits of epistemic democracy. *Soc Choice Welfare* 28(4): 649-666

[39] Wilson, R. (1975) On the Theory of Aggregation. *J Econ Theory* 10: 89-99

Franz Dietrich
Maastricht University and London School of Economics

Christian List
London School of Economics

Computing the nucleolus of weighted voting games

Edith Elkind and Dmitrii Pasechnik

Abstract

Weighted voting games (WVG) are coalitional games in which an agent's contribution to a coalition is given by his *weight*, and a coalition wins if its total weight meets or exceeds a given quota. These games model decision-making in political bodies as well as collaboration and surplus division in multiagent domains. The computational complexity of various solution concepts for weighted voting games received a lot of attention in recent years. In particular, Elkind et al.(2007) studied the complexity of stability-related solution concepts in WVGs, namely, of the core, the least core, and the nucleolus. While they have completely characterized the algorithmic complexity of the core and the least core, for the nucleolus they have only provided an NP-hardness result. In this paper, we solve an open problem posed by Elkind et al. by showing that the nucleolus of WVGs, and, more generally, k-vector weighted voting games with fixed k, can be computed in pseudopolynomial time, i.e., there exists an algorithm that correctly computes the nucleolus and runs in time polynomial in the number of players n and the maximum weight W. In doing so, we propose a general framework for computing the nucleolus, which may be applicable to a wider of class of games.

1 Introduction

Both in human societies and in multi-agent systems, there are many situations where individual agents can achieve their goals more efficiently (or at all) by working together. This type of scenarios is studied by *coalitional game theory*, which provides tools to decide which teams of agents will form and how they will divide the resulting profit. In general, to describe a coalitional game, one has to specify the payoff available to every team, i.e., every possible subset of agents. The size of such representation is exponential in the number of agents, and therefore working with a game given in such form is computationally intensive. For this reason, a lot of research effort has been spent on identifying and studying classes of coalitional games that correspond to rich and practically interesting classes of problems and yet have a compact representation.

One such class of coalitional games is *weighted voting games*, in which an agent's contribution to a coalition is given by his *weight*, and a coalition has value 1 if its total weight meets or exceeds a given quota, and 0 otherwise. These games model decision-making in political bodies, where agents correspond to political parties and the weight of each party is the number of its supporters, as well as task allocation in multi-agent systems, where the weight of each agent is the amount of resources it brings to the table and the quota is the total amount of resources needed to execute a task.

An important issue in coalitional games is *surplus division*, i.e., distributing the value of the resulting coalition between its members in a manner that encourages cooperation. In particular, it may be desirable that all agents work together, i.e., form the *grand coalition*. In this case, a natural goal is to distribute the payoff of the grand coalition so that it remains *stable*, i.e., so as to minimize the incentive for groups of agents to deviate and form coalitions of their own. Formally, this intuition is captured by several related solution concepts, such as the core, the least core, and the nucleolus. Without going into the technical details of their definitions (see Section 3), the nucleolus is, in some sense, the most stable payoff allocation scheme, and as such it is particularly desirable when the stability of the grand coalition is important.

The stability-related solution concepts for WVGs have been studied from computational perspective in [5]. There, the authors show that while computing the core is easy, finding the least core and the nucleolus is NP-hard. These computational hardness results rely on all weights being given in binary, which suggests that these problems may be easier for polynomially bounded weights. In-

deed, paper [5] provides a pseudopolynomial time algorithm (i.e., an algorithm whose running time is polynomial in the number of players n and the maximal weight W) for the least core. However, an analogous question for the nucleolus has been left open.

In this paper, we answer this question in affirmative by presenting a pseudopolynomial time algorithm for computing the nucleolus.

Theorem 1. *For a WVG specified by integer weights $w_1, \ldots, w_n$ and a quota q, there exists a procedure that computes its nucleolus in time polynomial in n and $W = \max_i w_i$.*

As in many practical scenarios (such as e.g., decision-making in political bodies) the weights are likely to be not too large, this provides a viable algorithmic solution to the problem of finding the nucleolus. Our approach relies on solving successive exponential-sized linear programs by constructing dynamic-programming based separation oracles, a technique that may prove useful in other applications.

A proof of Theorem 1 is presented in Section 4, after preliminaries in Section 3 and a discussion of related work in Section 2. The text rounds up by Section 5 that discusses conclusions and future work directions.

2 Related work

Another approach to payoff distribution in weighted voting games is based on *fairness*, i.e., dividing the payoff in a manner that is proportional to the agent's influence. The most popular solution concepts used in this context are the Shapley–Shubik power index [16] and the Banzhaf power index [1]. Both of these indices are known to be computationally hard for large weights [13, 4], yet efficiently computable for polynomially bounded weights [10].

The concept of the nucleolus was introduced by Schmeidler [14] in 1969. Paper [14] explains how the nucleolus arises naturally as "the most stable" payoff division scheme, and proves that the nucleolus is well-defined for any coalitional game and is unique. Kopelowitz [9] proposes to compute the nucleolus by solving a sequence of linear programs; we use this approach in our algorithm.

The computational complexity of the nucleolus has been studied for many classes of games, such as flow games [3], cyclic permutation games [17], assignment games [12], matching games [8], and neighbor games [7], as well as several others. While some of these papers provide polynomial-time algorithms for computing the nucleolus, others contain NP-hardness results.

The work in this paper is inspired by [5], which shows that the least core and the nucleolus of a weighted voting game are NP-hard to compute. It also proves that the nucleolus cannot be approximated within any constant factor. On the positive side, it provides a pseudopolynomial time algorithm for computing the least core, i.e., an algorithm whose running time is polynomial in n and W (rather than in the game representation size $O(n \log W)$), as well as a fully polynomial time approximation scheme (FPTAS) for the least core. However, for the nucleolus, paper [5] contains no algorithmic results.

3 Preliminaries and Notation

A *coalitional game* $G = (I, \nu)$ is given by a set of agents $I = \{1, \ldots, n\}$ and a function $\nu : 2^I \to \mathbb{R}$ that maps any subset (coalition) of the agents to a real value. This value is the total utility these agents can guarantee to themselves when working together. A coalitional game is called *simple* if $\nu(S) \in \{0, 1\}$ for any coalition $S \subseteq I$. In a simple game, a coalition S is called *winning* if $\nu(S) = 1$ and *losing* otherwise.

A *weighted voting game* is a simple coalitional game G given by a set of agents $I = \{1, \ldots, n\}$, their non-negative *weights* $\mathbf{w} = (w_1, \ldots, w_n)$, and a *quota* q; we write $G = (I; \mathbf{w}; q)$. As the focus

of this paper is computational complexity of such games, it is important to specify how the game is represented. In what follows, we assume that the weights and the quota are integers given in binary. This does not restrict the class of WVGs that we can work with, as any weighted voting game has such a representation [11].

For a coalition $S \subseteq I$, its value $\nu(S)$ is 1 (i.e., S is winning) if $\sum_{i \in S} w_i \geq q$; otherwise, $\nu(S) = 0$. Without loss of generality, we assume that the value of the grand coalition I is 1, that is, $\sum_{i \in I} w_i \geq q$. Also, we set $W = \max_{i \in I} w_i$.

For a coalitional game $G = (I, \nu)$, an *imputation* is a vector of non-negative numbers $\mathbf{p} = (p_1, \ldots, p_n)$, one for each agent in I, such that $\sum_{i \in I} p_i = \nu(I)$. We refer to p_i as the *payoff* of agent i. We write $p(S)$ to denote $\sum_{i \in S} p_i$. Similarly, $w(S)$ denotes $\sum_{i \in S} w_i$.

An important notion in coalitional games is that of stability: intuitively, a payoff vector should distribute the gains of the grand coalition in such a way that no group of agents has an incentive to deviate and form a coalition of their own. This intuition is captured by the notion of the core: the *core* of a game G is the set of all imputations $\mathbf{p}$ such that

$$p(S) \geq \nu(S) \text{ for all } S \subseteq I. \tag{1}$$

While the core is an appealing solution concept, it is very demanding: indeed, for many games of interest, the core is empty. In particular, it is well known that in simple games the core is empty unless there exists a veto player, i.e., a player that is present in all winning coalitions. Clearly, this is not always the case in weighted voting games, and a weaker solution concept is needed.

We can relax the notion of the core by allowing a small error in the inequalities (1). This leads to the notion of ε-core: the *ε-core* of a game G is the set of all imputations $\mathbf{p}$ such that $p(S) \geq \nu(S) - \varepsilon$ for all $S \subseteq I$. Under an imputation in the ε-core, the *deficit* of any coalition S, i.e., the difference $\nu(S) - p(S)$ between its value and the payoff that it gets, is at most ε. Observe that if ε is large enough, e.g., $\varepsilon \geq 1$, then the ε-core is guaranteed to be non-empty. Therefore, a natural goal is to identify the smallest value of ε such that the ε-core is non-empty, i.e., to minimize the error introduced by relaxing the inequalities in (1). This is captured by the concept of the least core, defined as the smallest non-empty ε-core of the game. More formally, consider the set $\{\varepsilon \mid \varepsilon \leq 1, \varepsilon\text{-core of } G \text{ is non-empty}\}$. It is easy to see that this set is compact, so it has a minimal element ε_1. The *least core* of G is its ε_1-core. The imputations in least core distribute the payoff in a way that minimizes the incentive to deviate: under any $\mathbf{p}$ in the least core, no coalition can gain more than ε_1 by deviating, and for any $\varepsilon' < \varepsilon_1$, there is no way to distribute the payoffs so that the deficit of every coalition is at most ε'. However, while the least core minimizes the worst-case deficit, it does not attempt to minimize the *number* of coalitions that experience the worst deficit, i.e., ε_1, nor does it try to minimize the second-worst deficit, etc. The *nucleolus* is a refinement of the least core that takes into account these higher-order effects.

Recall that the deficit of a coalition S under an imputation $\mathbf{p}$ is given by $d(\mathbf{p}, S) = \nu(S) - p(S)$. The *deficit vector* of $\mathbf{p}$ is the vector $\mathbf{d}(\mathbf{p}) = (d(\mathbf{p}, S_1), \ldots, d(\mathbf{p}, S_{2^n}))$, where $S_1, \ldots, S_{2^n}$ is a list of all subsets of I ordered so that $d(\mathbf{p}, S_1) \geq d(\mathbf{p}, S_2) \geq \cdots \geq d(\mathbf{p}, S_{2^n})$. In other words, the deficit vector lists the deficits of all coalitions from the largest to the smallest (which may be negative). The *nucleolus* is an imputation $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ that satisfies $\mathbf{d}(\boldsymbol{\eta}) \leq_{\text{lex}} \mathbf{d}(\mathbf{x})$ for any other imputation $\mathbf{x}$, where $\leq_{\text{lex}}$ is the lexicographic order. It is known [14] that the nucleolus is well-defined (i.e., an imputation with a lexicographically minimal deficit vector always exists) and is unique.

4 Algorithm

The description of our algorithm is structured as follows. We use the idea of [9], which explains how to compute the nucleolus by solving a sequence of (exponential-size) linear programs. In Section 4.1, we present the approach of [9], and argue that it correctly computes the nucleolus. This material is not new, and is presented here for completeness. In Section 4.2, we show how

to design separation oracles for the linear programs in this sequence so as to solve them by the ellipsoid method. While a naive implementation of these separation oracles would require storing exponentially many constraints, we show how to replace explicit enumeration of these constraints with a counting subroutine, while preserving the correctness of the algorithm. The arguments in Sections 4.1 and 4.2 apply to *any* coalitional game rather than just weighted voting games.

In Section 4.3, we show that for weighted voting games with polynomially-bounded weights the counting subroutine used by the algorithm of Section 4.2 can be efficiently implemented. Finally, in Section 4.4 we show how to modify this subroutine to efficiently identify a violated constraint if a given candidate solution is infeasible. The results in Sections 4.3 and 4.4 are specific to weighted voting games with polynomially bounded weights.

4.1 Computing the nucleolus by solving successive linear programs

As argued in [9], the nucleolus can be computed by solving at most n successive linear programs. The first linear program $\mathcal{LP}^1$ contains the inequality $p(S) \geq \nu(S) - \varepsilon$ for each coalition $S \subseteq I$, and attempts to minimize ε subject to these inequalities, i.e., it computes a payoff in the least core as well as the value ε^1 of the least core. Given a (*relative*) *interior optimizer* $(\mathbf{p}^1, \varepsilon^1)$ for $\mathcal{LP}^1$ (i.e., an optimal solution that minimizes the number of tight constraints), let Σ^1 be the set of all inequalities in $\mathcal{LP}^1$ that have been made tight by $\mathbf{p}^1$ (we will abuse notation and use Σ^1 to refer both to these inequalities and the corresponding coalitions). We construct the second linear program $\mathcal{LP}^2$ by replacing all inequalities in Σ^1 with equations of the form $p(S) = \nu(S) - \varepsilon^1$, and try to minimize ε subject to this new set of constraints. This results in $\varepsilon^2 < \varepsilon^1$ and a payoff vector $\mathbf{p}^2$ that satisfies $\mathbf{p}^2(S) = \nu(S) - \varepsilon^1$ for all $S \in \Sigma^1$, $\mathbf{p}^2(S) \geq \nu(S) - \varepsilon^2$ for all $S \notin \Sigma^1$. We repeat this process until the payoffs to all coalitions are determined, i.e., the solution space of the current linear program consists of a single point. It has been shown [9] that this will happen after at most n iterations: indeed, each iteration reduces the dimension of the solution space by at least 1.

More formally, the sequence of linear programs $(\mathcal{LP}^1, \ldots, \mathcal{LP}^n)$ is defined as follows. The first linear program $\mathcal{LP}^1$ is given by

$$\min_{(\mathbf{p},\varepsilon)} \varepsilon \quad \text{subject to} \quad \begin{cases} \sum_{i \in I} p_i = 1, \quad p_i \geq 0 \quad \text{for all } i = 1, \ldots, n \\ \sum_{i \in S} p_i \geq \nu(S) - \varepsilon \text{ for all } S \subseteq I. \end{cases} \tag{2}$$

Let $(\mathbf{p}^1, \varepsilon^1)$ be an interior optimizer to this linear program. Let Σ^1 be the set of tight constraints for $(\mathbf{p}^1, \varepsilon^1)$ (and, by a slight abuse of notation, the coalitions that correspond to them), i.e., for any $S \in \Sigma^1$ we have $p^1(S) = 1 - \varepsilon^1$.

Now, suppose that we have defined the first $j-1$ linear programs $\mathcal{LP}^1, \ldots, \mathcal{LP}^{j-1}$. For $k = 1, \ldots, j-1$, let $(\mathbf{p}^k, \varepsilon^k)$ be an interior optimizer for $\mathcal{LP}^k$ and let $\Sigma^k = \{S \mid p^k(S) = \nu(S) - \varepsilon^k\}$. Then the jth linear program $\mathcal{LP}^j$ is given by

$$\min_{(\mathbf{p},\varepsilon)} \varepsilon \quad \text{subject to} \quad \begin{cases} \sum_{i \in I} p_i = 1, \quad p_i \geq 0 \quad \text{for all } i = 1, \ldots, n \\ \sum_{i \in S} p_i = \nu(S) - \varepsilon^1 \text{ for all } S \in \Sigma^1 \\ \qquad \ldots \\ \sum_{i \in S} p_i = \nu(S) - \varepsilon^{j-1} \text{ for all } S \in \Sigma^{j-1} \\ \sum_{i \in S} p_i \geq \nu(S) - \varepsilon \quad \text{for all } S \notin \cup_{k=1}^{j-1} \Sigma^k. \end{cases} \tag{3}$$

Fix the minimal value of t such that there is no interior solution to $\mathcal{LP}^t$. It is not hard to see that the (unique) solution to $\mathcal{LP}^t$ is indeed the nucleolus. Indeed, the nucleolus is a payoff vector that produces the lexicographically maximal deficit vector. This means that it:

(i) minimizes ε^1 such that all coalitions receive at least $1 - \varepsilon^1$;

(ii) given (i), minimizes the number of coalitions that receive $1 - \varepsilon^1$;

(iii) given (i) and (ii), minimizes ε^2 such that all coalitions except for those receiving $1-\varepsilon^1$ receive at least $1 - \varepsilon^2$;

(iv) given (i), (ii) and (iii), minimizes the number of coalitions that receive $1 - \varepsilon^2$, etc.

Our sequence of linear programs finds a payoff vector that satisfies all these conditions; in particular, (ii) and (iv) (and analogous conditions at subsequent steps) are satisfied, since at each step we choose an interior optimizer for the corresponding linear program. The only issue that we have to address is that our procedure selects an *arbitrary* interior optimizer to the current linear program in order to construct the set Σ^j. Conceivably, this may have an impact on the final solution: if two interior optimizers to $\mathcal{LP}^j$ lead to two different sets Σ^j, they may also result in different values of ε^{j+1}, so one would have to worry about choosing the *right* interior optimizer. Fortunately, this is not the case, as shown by the following lemma.

Lemma 2. *Any two interior optimizers $(\mathbf{p}, \varepsilon)$ and $(\mathbf{q}, \varepsilon)$ for the linear program $\mathcal{LP}^j$ have the same set of tight constraints, i.e., the set Σ^j is independent of the choice of the interior optimizer.*

Proof. First note that the set of all interior optimizers for $\mathcal{LP}^j$ is convex. Now, suppose that $\mathbf{p}$ and $\mathbf{q}$ are two interior optimizers for $\mathcal{LP}^j$, but have different sets of tight constraints. Then, by convexity, any convex combination $\alpha\mathbf{p} + (1 - \alpha)\mathbf{q}$ of $\mathbf{p}$ and $\mathbf{q}$ is also an interior optimizer for $\mathcal{LP}^j$. However, the set of constraints that are tight for $\alpha\mathbf{p} + (1 - \alpha)\mathbf{q}$ is the intersection of the corresponding sets for $\mathbf{p}$ and $\mathbf{q}$, i.e., $\alpha\mathbf{p} + (1 - \alpha)\mathbf{q}$ has strictly fewer tight constraints than $\mathbf{p}$ or $\mathbf{q}$, a contradiction with $\mathbf{p}$ and $\mathbf{q}$ being interior optimizers for $\mathcal{LP}^j$. □

We conclude that when this algorithm terminates, the output is indeed the nucleolus. Next, we discuss how to solve each of the linear programs $\mathcal{LP}^j$, $j = 1, \ldots, t$.

4.2 Solving the linear programs $\mathcal{LP}^1, \ldots, \mathcal{LP}^t$

It is well-known (see e.g. [15, 6]) that a linear program can be solved in polynomial time by the ellipsoid method as long as it has a polynomial-time *separation oracle*, i.e., an algorithm that, given a candidate feasible solution, either confirms that it is feasible or outputs a violated constraint. Moreover, the ellipsoid method can also be used to find an interior optimizer (rather than an arbitrary optimal solution) in polynomial time [6, Thm. 6.5.5], as well as to decide whether one exists [6, Thm. 6.5.6]. We will now construct a polynomial-time separation oracle for jth linear program $\mathcal{LP}^j$ in our sequence.

It is easy to construct the part responsible for checking equations of $\mathcal{LP}^j$ in (3), assuming that we already have an oracle for the $(j-1)$st program $\mathcal{LP}^{j-1}$. Indeed, the latter oracle can be easily modified to also check whether the equation $\varepsilon = \varepsilon^{j-1}$ holds, thus providing an oracle for the optimal face of $\mathcal{LP}^{j-1}$. Then by [6, Thm. 6.5.5] we can compute a basis of the optimal face (which consists of at most n equations) in polynomial time. The separation oracle can then reject a candidate solution $(\mathbf{p}, \varepsilon)$ if $\mathbf{p}$ violates one of those basis equations.

Dealing with the inequalities of $\mathcal{LP}^j$ is more complicated. A naive separation oracle would have to explicitly list the sets $\Sigma^1, \ldots, \Sigma^{j-1}$, which may be superpolynomial in size. Alternatively, one can treat this part of the oracle as a 0-1 integer linear feasibility problem, with 0-1 variables x_i encoding a set $S \notin \cup_{k=1}^{j-1}\Sigma^k$ that provides a separating inequality for the oracle input $(\mathbf{p}, \varepsilon)$.

Namely, suppose that we have verified that $\mathbf{p}$ satisfies all the equations in $\mathcal{LP}^j$ (as described above). Then, given an interior optimizer $(\mathbf{p}^{j-1}, \varepsilon^{j-1})$ for $\mathcal{LP}^{j-1}$, the values $x_1, \ldots, x_n$ can be obtained as a solution to the following inequalities:

$$\sum_i p_i^{j-1} x_i > 1 - \varepsilon^{j-1}, \tag{4}$$

$$\sum_i p_i x_i < 1 - \varepsilon, \tag{5}$$

$$\sum_i w_i x_i \geq q. \tag{6}$$

The problem with this approach is that for arbitrary rational $\mathbf{p}$ and $\mathbf{p}^{j-1}$ this system of inequalities is at least as hard as KNAPSACK, which is NP-complete. Moreover, as $\mathbf{p}$ and $\mathbf{p}^{j-1}$ are produced by the ellipsoid method, there is no guarantee that their bitsizes are small enough to use a (pseudo)polynomial-time algorithm for KNAPSACK. The only hope is to replace at least one of (4) and (5) by something "tame".

We will now present a more sophisticated approach to identifying a violated constraint. In a way, it can be seen as replacing checking (4) with counting. Our construction proceeds by induction: to construct a separation oracle for $\mathcal{LP}^j$, we assume that we have constructed an oracle for $\mathcal{LP}^{j-1}$, and are given the sizes of sets $\Sigma^1, \ldots, \Sigma^{j-1}$ as well as the sequence $(\varepsilon^1, \ldots, \varepsilon^{j-1})$.

By construction, any optimal solution $(\mathbf{p}, \varepsilon)$ to $\mathcal{LP}^j$ satisfies $\varepsilon < \varepsilon^{j-1}$, so we can add the constraint $\varepsilon \leq \varepsilon^{j-1}$ to $\mathcal{LP}^j$ without changing the set of solutions. From now on, we will assume that $\mathcal{LP}^j$ includes this constraint. Our separation oracle will first check whether a given candidate solution $(\mathbf{p}, \varepsilon)$ satisfies $\varepsilon \leq \varepsilon^{j-1}$, as well as constraints $p_i \geq 0$ for all $i = 1, \ldots, n$ and $p(I) = 1$, and reject $(\mathbf{p}, \varepsilon)$ and output a violated constraint if this is not the case. Therefore, in what follows we assume that $(\mathbf{p}, \varepsilon)$ satisfies all these easy-to-identify constraints.

Now, a candidate solution $(\mathbf{p}, \varepsilon)$ is feasible for $\mathcal{LP}^j$ if $p(S) = \nu(S) - \varepsilon^t$ for $S \in \Sigma^t$, $t = 1, \ldots, j-1$, and $p(S) \geq \nu(S) - \varepsilon$ for all $S \notin \cup_{t=1}^{j-1} \Sigma^t$. Recall that the deficit of a coalition $S \subseteq I$ under a payoff vector $\mathbf{p}$ is given by $\nu(S) - p(S)$. Suppose that we have a procedure $\mathcal{P}(\mathbf{p}, \varepsilon)$ that, given a candidate solution $(\mathbf{p}, \varepsilon)$, can efficiently compute the top j distinct deficits under $\mathbf{p}$, i.e.,

$$\begin{aligned}
m^1 &= \max\{d(S) \mid S \subseteq I\} \\
m^2 &= \max\{d(S) \mid S \subseteq I, d(S) \neq m^1\} \\
&\ldots \\
m^j &= \max\{d(S) \mid S \subseteq I, d(S) \neq m^1, \ldots, m^{j-1}\}
\end{aligned}$$

as well as the numbers $n^1, \ldots, n^j$ of coalitions that have deficits of $m^1, \ldots, m^j$, respectively:

$$n^k = |\{S \mid S \subseteq I, d(S) = m^k\}|, \qquad k = 1, \ldots, j.$$

Suppose also that we are given the values $\varepsilon^1, \ldots, \varepsilon^{j-1}$ and the sizes s^t of the sets Σ^t, $t = 1, \ldots, j-1$.

Now, our algorithm works as follows. Given a candidate solution $(\mathbf{p}, \varepsilon)$, it runs $\mathcal{P}(\mathbf{p}, \varepsilon)$ to obtain m^t, n^t, $t = 1, \ldots, j$. If $\varepsilon < \varepsilon^{j-1}$, it then checks whether

(a) $m^t = \varepsilon^t$ and $n^t = s^t$ for all $t = 1, \ldots, j-1$

(b) $m^j \leq \varepsilon$.

If $\varepsilon = \varepsilon^{j-1}$, it simply checks whether $m^t = \varepsilon^t$ for $t = 1, \ldots, j-1$ and $n^t = s^t$ for all $t = 1, \ldots, j-2$ [1]. If these conditions are satisfied, the algorithm answers that $(\mathbf{p}, \varepsilon)$ is indeed a feasible

[1] Alternatively, if $\varepsilon = \varepsilon^j$, one can verify whether $(\mathbf{p}, \varepsilon)$ is a feasible solution to the previous linear program $\mathcal{LP}^{j-1}$.

solution, and otherwise it identifies and outputs a violated constraint (for details of this step, see Section 4.4). We will now show that this algorithm implements a separation oracle for $\mathcal{LP}^j$ correctly and efficiently.

Theorem 3. *Given the values ε^t, s^t, $t = 1, \ldots, j-1$, and a procedure $\mathcal{P}(\mathbf{p}, \varepsilon)$ that computes m^t, n^t, $t = 1, \ldots, j$, in polynomial time, our algorithm correctly decides whether a given pair $(\mathbf{p}, \varepsilon)$ is feasible for $\mathcal{LP}^j$ and runs in polynomial time.*

Proof. We start by proving an auxiliary lemma.

Lemma 4. *For any vector $\mathbf{p}$ and any $t \leq j-1$ such that $m^s = \varepsilon^s$, $n^s = s^s$ for all $s \leq t$, the coalitions with deficit ε^s under $\mathbf{p}$ are exactly the ones in Σ^s.*

Proof. The proof is by induction on s. For $s = 1$, we have that the largest deficit of any coalition under $\mathbf{p}$ is ε^1, and there are exactly s^1 coalitions with this deficit. Hence, $(\mathbf{p}, \varepsilon^1)$ is an interior optimizer for $\mathcal{LP}^1$, and therefore the lemma follows by Lemma 2. Now, suppose that the lemma has been proven for $s-1$. By the induction hypothesis, $\mathbf{p}$ satisfies all constraints in $\Sigma^1, \ldots, \Sigma^{s-1}$. Also, under $\mathbf{p}$ there are at most $s^1 + \cdots + s^{s-1}$ coalitions whose deficit exceeds ε^s, so for all coalitions not in $\cup_{r=1}^{s-1}\Sigma^r$ their deficit is at most ε^s. Finally, there are exactly s^s coalitions whose deficit is exactly ε^s. Hence, $(\mathbf{p}, \varepsilon^s)$ is an interior optimizer for $\mathcal{LP}^s$, and therefore the lemma follows by Lemma 2. □

To prove the theorem, let us first consider the case $\varepsilon < \varepsilon^{j-1}$. Suppose that $(\mathbf{p}, \varepsilon)$ satisfies (a) and (b). By using Lemma 4 with $t = j-1$, we conclude that $(\mathbf{p}, \varepsilon)$ satisfies all equations in $\mathcal{LP}^j$. Now, under $\mathbf{p}$, m^j is the largest deficit of a coalition not in $\cup_{t=1}^{j-1}\Sigma^t$. If this deficit is at most ε, then the pair $(\mathbf{p}, \varepsilon)$ is a feasible solution to $\mathcal{LP}^j$.

Conversely, suppose that (a) or (b) is violated. If $(\mathbf{p}, \varepsilon)$ satisfies (a) but violates (b), by using Lemma 4 with $t = j-1$ we conclude that, under $\mathbf{p}$ the coalitions in Σ^t have deficit ε^t for $t = 1, \ldots, j-1$, but the deficit of some coalition not in $\cup_{t=1}^{j-1}\Sigma^t$ exceeds ε. Hence, this coalition corresponds to a violated constraint. Now, suppose that (a) does not hold, and let s be the smallest index for which $m^s \neq \varepsilon^s$ or $n^s \neq s^s$. By using Lemma 4 with $t = s-1$, we conclude that for $r = 1, \ldots, s-1$ the coalitions in Σ^r have deficit ε^r. However, either the sth distinct deficit under $\mathbf{p}$ is not ε^s, in which case $\mathbf{p}$ violates a constraint in $2^I \setminus \cup_{r=1}^{s-1}\Sigma^r$, or under $\mathbf{p}$ there are more than s^s coalitions with deficit ε^s (note that by construction it cannot be the case that $n^s < s^s$). In the latter case, there is a coalition in $2^I \setminus \cup_{r=1}^{s}\Sigma^r$ whose deficit exceeds ε^s, thus violating the corresponding constraint. The case $\varepsilon = \varepsilon^{j-1}$ is similar. In this case for a candidate solution $(\mathbf{p}, \varepsilon)$ to be feasible, it is not required that there are exactly s^{j-1} coalitions with deficit ε^{j-1}. Hence, the algorithm only has to decide if for all $t = 1, \ldots, j-2$, the coalitions with deficit ε^t under $\mathbf{p}$ are exactly the ones in Σ^t, and all other coalitions get at least ε^{j-1}. Showing that our algorithm checks this correctly can be done similarly to the previous case.

The bound on the running time is obvious from the description of the algorithm. □

To provide the value $s^j = |\Sigma^j|$ for the subsequent linear programs $\mathcal{LP}^{j+1}, \ldots, \mathcal{LP}^n$, we need to find an interior optimizer for $\mathcal{LP}^j$. Thm. 6.5.5 in [6] explains how to do this given a separation oracle for the optimal face, i.e., the set of all optimizers of $\mathcal{LP}^j$. Observe that such an oracle can be obtained by a slight modification of the oracle described above. Indeed, the optimal face is the set of solutions to the linear feasibility problem given by the constraints in $\mathcal{LP}^j$ together with the constraint $\varepsilon = \varepsilon^j$. The modified oracle first checks the latter constraint, reports the violation (and the corresponding inequality) if it happens, and otherwise continues as the original oracle. Clearly, the modified oracle runs in polynomial time whenever the original one does. Hence, we can compute s^j in polynomial time by computing an interior solution $(\mathbf{p}, \varepsilon)$ to $\mathcal{LP}^j$ according to [6, Thm. 6.5.5], running $\mathcal{P}(\mathbf{p}, \varepsilon)$ to find n^j, and setting $s^j = n^j$.

4.3 Implementing the counting

We will now show how to implement the counting procedure $\mathcal{P}(\mathbf{p}, \varepsilon)$ used in Section 4.2 for WVGs. The running time of our procedure is polynomial in the number of players n and the maximum weight W.

Our approach is based on dynamic programming. Fix a WVG $(I; \mathbf{w}; q)$, a payoff vector $\mathbf{p}$, and $j \leq n$. For all $k = 1, \ldots, n$, $w = 1, \ldots, nW$, let $X^1_{k,w}, \ldots, X^j_{k,w}$ be the bottom j distinct payoffs to coalitions in $\{1, \ldots, k\}$ of weight w, i.e., define

$$\begin{aligned}
X^1_{k,w} &= \min\{p(S) \mid S \subseteq \{1, \ldots, k\}, w(S) = w\} \\
X^2_{k,w} &= \min\{p(S) \mid S \subseteq \{1, \ldots, k\}, w(S) = w, p(S) \neq X^1_{k,w}\} \\
&\ldots \\
X^j_{k,w} &= \min\{p(S) \mid S \subseteq \{1, \ldots, k\}, w(S) = w, p(S) \neq X^1_{k,w}, \ldots, X^{j-1}_{k,w}\}
\end{aligned}$$

and let $Y^1_{k,w}, \ldots, Y^j_{k,w}$ be the numbers of coalitions that get these payoffs, i.e. set

$$Y^t_{k,w} = |\{S \mid S \subseteq \{1, \ldots, k\}, w(S) = w, p(S) = X^t_{k,w}\}|, \qquad \text{for } t = 1, \ldots, j.$$

These quantities can be computed inductively for $k = 1, \ldots, n$ as follows.

For $k = 1$, we have $X^1_{1,w} = p_1$ if $w = w_1$ and $+\infty$ otherwise, $Y^1_{1,w} = 1$ if $w = w_1$ and 0 otherwise, and $X^t_{1,w} = +\infty$, $Y^t_{1,w} = 0$ for $t = 2, \ldots, j$.

Now, suppose that we have computed $X^1_{k-1,w}, \ldots, X^j_{k-1,w}, Y^1_{k-1,w}, \ldots, Y^j_{k-1,w}$ for all $w = 1, \ldots, nW$. Consider $S \subseteq \{1, \ldots, k\}$ receiving one of the bottom j distinct payoffs to subsets of $\{1, \ldots, k\}$ of weight w, i.e., $p(S) \in \{X^1_{k,w}, \ldots, X^j_{k,w}\}$. Then either

(1) $S \subseteq \{1, \ldots, k-1\}$, in which case S must be among the coalitions that receive one of the bottom j distinct payoffs to subsets of $\{1, \ldots, k-1\}$ of weight w, i.e., we have $p(S) \in \{X^1_{k-1,w}, \ldots, X^j_{k-1,w}\}$, or

(2) $k \in S$, in which case $S \setminus \{k\}$ must be among the coalitions that receive one of the bottom j distinct payoffs to subsets of $\{1, \ldots, k-1\}$ of weight $w - w_k$, i.e., we have $p(S \setminus \{k\}) \in \{X^1_{k-1,w-w_k}, \ldots, X^j_{k-1,w-w_k}\}$.

Consider the multi-set $\mathcal{S}_{k,w} = \{X^1_{k-1,w}, \ldots, X^j_{k-1,w}, p_k + X^1_{k-1,w-w_k}, \ldots, p_k + X^j_{k-1,w-w_k}\}$. By the argument above, we have

$$\begin{aligned}
X^1_{k,w} &= \min\{x \mid x \in \mathcal{S}_{k,w}\} \\
X^2_{k,w} &= \min\{x \mid x \in \mathcal{S}_{k,w}, x \neq X^1_{k,w}\} \\
&\ldots \\
X^j_{k,w} &= \min\{x \mid x \in \mathcal{S}_{k,w}, x \neq X^1_{k,w}, \ldots, X^{j-1}_{k,w}\}.
\end{aligned}$$

The number of coalitions that receive the payoff $X^t_{k,w}$, i.e., $Y^t_{k,w}$, $t = 1, \ldots, j$, depends on how many times $X^t_{k,w}$ appears in $\mathcal{S}_{k,w}$. If it only appears once, then there is only one source of sets that receive a payoff of $X^t_{k,w}$, i.e., we set $Y^t_{k,w} = Y^s_{k-1,w}$ if $X^t_{k,w}$ appears as $X^s_{k-1,w}$ for some $s = 1, \ldots, j$, and we set $Y^t_{k,w} = Y^s_{k-1,w-w_k}$ if $X^t_{k,w}$ appears as $X^s_{k-1,w-w_k} + p_k$ for some $s = 1, \ldots, j$. On the other hand, if $X^t_{k,w}$ appears twice in $\mathcal{S}_{k,w}$ (first time as $X^s_{k-1,w}$ and second time as $p_k + X^{s'}_{k-1,w-w_k}$ for some $s, s' = 1, \ldots, j$), we have to add up the corresponding counts, i.e., we set $Y^t_{k,w} = Y^s_{k-1,w} + Y^{s'}_{k-1,w-w_k}$.

After all $X^1_{n,w}, \ldots, X^j_{n,w}, Y^1_{n,w}, \ldots, Y^j_{n,w}$ have been evaluated, it is not hard to compute m^t, n^t, $t = 1, \ldots, j$. Indeed, the top j deficits appear in the multi-set

$$\mathcal{S} = \{I_w - X^1_{n,w}, \ldots, I_w - X^j_{n,w} \mid w = 1, \ldots, nW\},$$

where $I_w = 1$ if $w \geq q$ and $I_w = 0$ if $w < q$ (recall that q is the quota of the game, i.e., $\nu(S) = 1$ if and only if $w(S) \geq q$). Hence, we can set

$$\begin{aligned} m^1 &= \max\{x \mid x \in \mathcal{S}\} \\ m^2 &= \max\{x \mid x \in \mathcal{S}, x \neq m^1\} \\ &\cdots \\ m^j &= \max\{x \mid x \in \mathcal{S}, x \neq m^1, \ldots, m^{j-1}\}. \end{aligned}$$

The procedure for computing n^t, $t = 1, \ldots, j$, is similar to that of computing $Y^s_{k,w}$ (see above): we have to check how many times m^t appears in $\mathcal{S}$ and add the corresponding counts.

In the next subsection, we will show how to find a violated inequality if $(\mathbf{p}, \varepsilon)$ is not a feasible solution to $\mathcal{LP}^j$.

4.4 Identifying a violated constraint

Consider $\mathcal{LP}^j$ and a candidate solution $(\mathbf{p}, \varepsilon)$. Suppose that the algorithm described in the previous subsection has decided that $(\mathbf{p}, \varepsilon)$ is not a feasible solution to $\mathcal{LP}^j$. This can happen in three possible ways.

(a) $m^s = \varepsilon^s$, $n^s = s^s$ for $s = 1, \ldots, \ell - 1$, but $m^\ell \neq \varepsilon^\ell$ for an $\ell < j$.

(b) $m^s = \varepsilon^s$, $n^s = s^s$ for $s = 1, \ldots, \ell - 1$, $m^\ell = \varepsilon^\ell$, but $n^\ell \neq s^\ell$ for an $\ell < j$.

(c) $m^s = \varepsilon^s$, $n^s = s^s$ for $s = 1, \ldots, j - 1$, but $m^j > \varepsilon$.

In cases (a) and (b), there is a violated equation in (3), while in (c) there are none (but there is a violated inequality). Thus (a) and (b) can be handled using the ideas discussed in the beginning of Section 4.2. Indeed, as argued there, we can efficiently compute the basis of the optimal face of the feasible set of $\mathcal{LP}^{j-1}$ using the ellipsoid method. One can then easily check if a candidate solution violates one of the equations in the basis (recall that there are at most n of them), and, if this is the case, report one that is violated. Hence, we only need to show how to identify a violated constraint in case (c). However, for completeness, we present here a purely counting-based algorithm for each of the cases.

In case (a), let $(\hat{\mathbf{p}}, \varepsilon^\ell)$ be an interior optimizer for $\mathcal{LP}^\ell$. Under $\hat{\mathbf{p}}$, the deficit of any coalition in $2^I \setminus \cup_{s=1}^{\ell-1}\Sigma^s$ is at most ε^ℓ. On the other hand, under $\mathbf{p}$, there are n^ℓ coalitions in $2^I \setminus \cup_{s=1}^{\ell-1}\Sigma^s$ whose deficit is $m^\ell > \varepsilon^\ell$. Each of these coalitions corresponds to a violated constraint: indeed, if such a coalition is in Σ^s, $s \geq \ell$, then $\mathcal{LP}^j$ requires that its deficit is $\varepsilon^s \leq \varepsilon^\ell < m^\ell$, and if it is in $2^I \setminus \cup_{s=1}^{j-1}\Sigma^s$, then $\mathcal{LP}^j$ requires that its deficit is at most $\varepsilon \leq \varepsilon^\ell < m^\ell$, Hence, it suffices to identify a coalition whose deficit under $\mathbf{p}$ is m^ℓ. To this end, we can modify the dynamic program for $\mathbf{p}$ as follows. Together with every variable $X^t_{k,w}$, $t = 1, \ldots, j$, $k = 1, \ldots, n$, $w = 1, \ldots, nW$, we will use an auxiliary variable $Z^t_{k,w}$ which stores a coalition whose payoff under $\mathbf{p}$ is equal to $X^t_{k,w}$. The values of $Z^t_{k,w}$ can be easily computed by induction: if $X^t_{k,w} = X^s_{k-1,w}$ for some $s = 1, \ldots, j$ then $Z^t_{k,w} = Z^s_{k-1,w}$, and if $X^t_{k,w} = p_k + X^s_{k-1,w-w_k}$ for some $s = 1, \ldots, j$ then $Z^t_{k,w} = Z^s_{k-1,w} \cup \{k\}$ (if $X^t_{k,w} = X^s_{k-1,w} = p_k + X^{s'}_{k-1,w-w_k}$, we can set $Z^t_{k,w}$ to either of these values). Now, there exist some t and w such that $w \geq q$ and $1 - X^t_{n,w} = m^\ell$ or $w < q$ and $-X^t_{n,w} = m^\ell$; such t and w can be found by scanning all $X^t_{n,w}$. The corresponding set $Z^t_{n,w}$ has deficit m^ℓ under $\mathbf{p}$, and therefore corresponds to a violated constraint.

In case (b), as before, let $(\mathbf{p}, \varepsilon)$ be an interior optimizer to $\mathcal{LP}^{j-1}$. There exists a coalition whose deficit under $\mathbf{p}$ is ε^ℓ, but whose deficit under $\hat{\mathbf{p}}$ is strictly less than ε^ℓ. To find such a coalition, run $\mathcal{P}(\hat{\mathbf{p}}, \varepsilon)$ in order to compute the corresponding values $\hat{X}^t_{k,w}$, $\hat{Y}^t_{k,w}$ $t = 1, \ldots, j-1$, $k = 1, \ldots, n$, $w = 1, \ldots, nW$. Define Z_w as follows: if there exists some $t \in \{1, \ldots, j-1\}$ such that $X^t_{n,w} =$

$I_w - \varepsilon^\ell$, set $Z_w = Y^t_{n,w}$; otherwise, set $Z_w = 0$. $\hat{Z}_w$ is defined similarly: if there exists an $s \in \{1, \ldots, j-1\}$ such that $\hat{X}^s_{n,w} = I_w - \varepsilon^\ell$, set $\hat{Z}_w = \hat{Y}^s_{n,w}$; otherwise, set $\hat{Z}_w = 0$. The variables Z_w and $\hat{Z}_w$ count the number of coalitions that have total weight w and have deficit ε^ℓ under $\mathbf{p}$ and $\hat{\mathbf{p}}$, respectively. We have $n^\ell = \sum_{w=1,\ldots,nW} Z_w$, $s^\ell = \sum_{w=1,\ldots,nW} \hat{Z}_w$. As $n^\ell > s^\ell$, there exists a weight w such that $Z_w > \hat{Z}_w$. Set $q = I_w - \varepsilon^\ell$, i.e., q is the total payment received by the coalitions counted by Z_w and $\hat{Z}_w$. Now, we have $Z_w = Z^n_w + Z^{-n}_w$, where Z^n_w is the number of coalitions of weight w that include n, have weight w and receive total payment q, and Z^{-n}_w is the number of coalitions of weight w that do not include n, have weight w and receive total payment ε^ℓ; $\hat{Z}^n_w$ and $\hat{Z}^{-n}_w$ can be defined similarly. We can easily compute these quantities: for example, Z^n_w is the number of subsets of $\{1, \ldots, n-1\}$ that have weight $w - w_n$ and receive total payment $q - p_n$, i.e. $Z^n_w = Y^t_{n-1,w-w_n}$ if there exists a $t \in \{1, \ldots, j-1\}$ such that $X^t_{n-1,w-w_n} = q - p_n$, and $Z^{-n}_w = 0$ otherwise. It follows immediately that $Z^n_w > \hat{Z}^n_w$ or $Z^{-n}_w > \hat{Z}^{-n}_w$ (or both), and we can easily verify which of these cases holds. In the former case, we can conclude that the number of coalitions in $\{1, \ldots, n-1\}$ that have weight $w - w_n$ and are paid $q - p_n$ under $\mathbf{p}$ exceeds the number of coalitions in $\{1, \ldots, n-1\}$ that have weight $w - w_n$ and are paid $q - \hat{p}_n$ under $\hat{\mathbf{p}}$. In the latter case, we can conclude that the number of coalitions in $\{1, \ldots, n-1\}$ that have weight w and are paid q under $\mathbf{p}$ exceeds the number of coalitions in $\{1, \ldots, n-1\}$ that have weight w and are paid q under $\hat{\mathbf{p}}$. Continuing in the same manner for $n-1, \ldots, 1$, we can identify a coalition that is paid q under $\mathbf{p}$, but not under $\hat{p}$.

Case (c), i.e. $m^j > \varepsilon$, is similar to (a) and can be handled in the same manner.

5 Conclusions and future work

In this paper, we proposed a new technique for computing the nucleolus of coalitional games. Namely, we have shown that, when constructing the separation oracle for the jth linear program $\mathcal{LP}^j$, instead of storing the sets of tight constraints for the linear programs $\mathcal{LP}^t$, $t = 1, \ldots, j-1$, it suffices to store the sizes of these sets as well as the top $j-1$ deficits of an interior optimizer $(\mathbf{p}^{j-1}, \varepsilon)$ to $\mathcal{LP}^{j-1}$. A feasibility of a candidate solution $(\mathbf{p}, \varepsilon)$ to $\mathcal{LP}^j$ can then be verified, roughly, by computing the top j deficits for $\mathbf{p}$ as well as the number of coalitions that have these deficits, and comparing these values to their pre-computed counterparts for $(\mathbf{p}^{j-1}, \varepsilon)$.

We then demonstrated the usefulness of this technique by showing that for weighted voting games with polynomially-bounded weights both the top j deficits and the number of coalitions that have these deficits can be efficiently computed using dynamic programming. This allows us to implement the separation oracles for our linear programs in pseudopolynomial time. Combining this with the ellipsoid algorithm results in a pseudopolynomial time algorithm for the nucleolus of weighted voting games, thus solving an open problem posed by [5]. Furthermore, the general technique put forward in this paper effectively reduces the computation of the nucleolus to solving a natural combinatorial problem for the underlying game. Namely, we can state the following meta-theorem:

Theorem 5. *Given a coalitional game G, suppose that we can, for any payoff vector $\mathbf{p}$, identify the top n distinct deficits under $\mathbf{p}$ as well as the number of coalitions that have these deficits in polynomial time. Then we can compute the nucleolus of G in polynomial time.*

We believe that this framework can be useful for computing the nucleolus in other classes of games. Indeed, by stripping away most of the game-theoretic terminology, we may be able to find the nucleolus by applying existing results in combinatorics and discrete mathematics in a black-box fashion.

In the context of weighted voting games, our assumption that the weights are polynomially bounded (or, equivalently, given in unary) is essential, as [5] shows that the nucleolus is NP-hard to

compute for WVGs with weights given in binary. Moreover, in many practical scenarios the agents' weights cannot be too large (e.g., polynomial functions of n), in which case the running time of our algorithm is polynomial.

By a slight modification of our algorithm, we can obtain a pseudopolynomial time algorithm for computing the nucleolus in *k-vector weighted voting games* for constant k. Informally speaking, these are games given by the intersection of k weighted voting games, i.e., a coalition is considered to be winning if it wins in each of the underlying games. There are some interesting games that can be represented as k-vector weighted voting games for small values of k (i.e., $k = 2$ or $k = 3$), but not as weighted voting games, most notably, voting in the European Union [2]. Hence, this extension of our algorithm enables us to compute the nucleolus in some real-life scenarios. The overall structure of our algorithm remains the same. However, the dynamic program has to be modified to keep track of several weight systems simultaneously.

Another natural way to address the problem of computing the nucleolus is by focusing on approximate solutions. Indeed, [5] proposes a fully polynomial time approximation scheme (FPTAS) for several least-core related problems. It would be natural to expect a similar result to hold for the nucleolus. Unfortunately, this approach is ruled out by [5], which shows that it is NP-hard to decide whether the nucleolus payoff of any particular player is 0, and therefore approximating the nucleolus payoffs up to any constant factor is NP-hard. Nevertheless, one can attempt to find an *additive* approximation to the nucleolus, i.e., for a given error bound $\delta > 0$, find a vector $\mathbf{x}$ such that $|\eta_i - x_i| \leq \delta$ for $i = 1, \ldots, n$. This can be useful in situations when the agents' weights cannot be assumed to be polynomially bounded with respect to n, e.g., in the multiagent settings where the weights correspond to agents' resources. We are currently investigating several approaches to designing additive approximation algorithms for the nucleolus.

References

[1] J. F. Banzhaf. Weighted voting doesn't work: a mathematical analysis. *Rutgers Law Review*, 19:317–343, 1965.

[2] J. M. Bilbao, J. R. Fernández, N. Jiminéz, and J. J. López. Voting power in the European Union enlargement. *European Journal of Operational Research*, 143:181–196, 2002.

[3] X. Deng, Q. Fang, and X. Sun. Finding nucleolus of flow game. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 124–131, New York, 2006. ACM.

[4] X. Deng and C. H. Papadimitriou. On the complexity of cooperative solution concepts. *Mathematics of Operations Research*, 19(2):257–266, 1994.

[5] E. Elkind, L. A. Goldberg, P. W. Goldberg, and M. Wooldridge. Computational complexity of weighted threshold games. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*, volume 1, pages 718–723, Menlo Park, California, 2007. AAAI Press. ISBN 978-1-57735-323-2.

[6] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2 of *Algorithms and Combinatorics*. Springer-Verlag, Berlin, second edition, 1993.

[7] H. Hamers, F. Klijn, T. Solymosi, S. Tijs, and D. Vermeulen. On the nucleolus of neighbor games. *European J. Oper. Res.*, 146(1):1–18, 2003.

[8] W. Kern and D. Paulusma. Matching games: the least core and the nucleolus. *Math. Oper. Res.*, 28(2):294–308, 2003.

[9] A. Kopelowitz. Computation of the kernels of simple games and the nucleolus of n-person games. Preprint RM 37, 1967. Research Program in Game Theory and Mathematical Economics.

[10] T. Matsui and Y. Matsui. A survey of algorithms for calculating power indices of weighted majority games. *J. Oper. Res. Soc. Japan*, 43(1):71–86, 2000. New trends in mathematical programming (Kyoto, 1998).

[11] S. Muroga. *Threshold Logic and its Applications*. John Wiley & Sons, 1971.

[12] M. Núñez. A note on the nucleolus and the kernel of the assignment game. *Internat. J. Game Theory*, 33(1):55–65, 2004.

[13] K. Prasad and J. S. Kelly. NP-completeness of some problems concerning voting games. *Internat. J. Game Theory*, 19(1):1–9, 1990.

[14] D. Schmeidler. The nucleolus of a characteristic function game. *SIAM J. Appl. Math.*, 17:1163–1170, 1969.

[15] A. Schrijver. *Theory of linear and integer programming*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons Ltd., Chichester, 1986. A Wiley-Interscience Publication.

[16] L. S. Shapley and M. Shubik. A method for evaluating the distribution of power in a committee system. In *The Shapley value*, pages 41–48. Cambridge Univ. Press, Cambridge, 1988.

[17] T. Solymosi, T. E. S. Raghavan, and S. Tijs. Computing the nucleolus of cyclic permutation games. *European J. Oper. Res.*, 162(1):270–280, 2005.

Edith Elkind
Intelligence, Agents, Multimedia group
School of Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ, United Kingdom
Email: ee@ecs.soton.ac.uk

Dmitrii Pasechnik
Division of Mathematical Sciences
School of Physical and Mathematical Sciences
Nanyang Technological University
21 Nanyang Link
Singapore 637371
Email: dima@ntu.edu.sg

Sincere-Strategy Preference-Based Approval Voting Fully Resists Constructive Control and Broadly Resists Destructive Control[1]

Gábor Erdélyi, Markus Nowak, and Jörg Rothe

Abstract

We study sincere-strategy preference-based approval voting (SP-AV), a system proposed by Brams and Sanver [8], with respect to procedural control. In such control scenarios, an external agent seeks to change the outcome of an election via actions such as adding/deleting/partitioning either candidates or voters. SP-AV combines the voters' preference rankings with their approvals of candidates, and we adapt it here so as to keep its useful features with respect to approval strategies even in the presence of control actions. We prove that this system is computationally resistant (i.e., the corresponding control problems are NP-hard) to 19 out of 22 types of constructive and destructive control. Thus, SP-AV has more resistances to control, by three, than is currently known for any other natural voting system with a polynomial-time winner problem. In particular, SP-AV is (after Copeland voting, see [19]) the second natural voting system with an easy winner-determination procedure that is known to have full resistance to constructive control, and unlike Copeland voting it in addition displays broad resistance to destructive control.

1 Introduction

Voting provides a particularly useful method for preference aggregation and collective decision-making. While voting systems were originally used in political science, economics, and operations research, they are now also of central importance in various areas of computer science, such as artificial intelligence (in particular, within multiagent systems). In automated, large-scale computer settings, voting systems have been applied, e.g., for planning [11] and similarity search [15], and have also been used in the design of recommender systems [21] and ranking algorithms [10] (where they help to lessen the spam in meta-search web-page rankings). For such applications, it is crucial to explore the computational properties of voting systems and, in particular, to study the complexity of problems related to voting (see, e.g., the survey by Faliszewski et al. [17]).

The study of voting systems from a complexity-theoretic perspective was initiated by Bartholdi, Tovey, and Trick's series of seminal papers about the complexity of winner determination [2], manipulation [1], and procedural control [3] in elections. This paper contributes to the study of electoral control, where an external agent—traditionally called *the chair*—seeks to influence the outcome of an election via procedural changes to the election's structure, namely via adding/deleting/partitioning either candidates or voters (see Section 2.2 for the formal definitions of our control problems). We consider both *constructive* control (introduced by Bartholdi et al. [3]), where the chair's goal is to make a given candidate the unique winner, and *destructive* control (introduced by Hemaspaandra et al. [22]), where the chair's goal is to prevent a given candidate from being a unique winner.

[1] Supported in part by the DFG under grants RO 1202/12-1 (within the European Science Foundation's EUROCORES program LogICCC: "Computational Foundations of Social Choice") and RO 1202/11-1 and by the Alexander von Humboldt Foundation's TransCoop program. Some results of this paper were presented at the 33rd International Symposium on Mathematical Foundations of Computer Science (MFCS-08), August 2008 [13].

We investigate the same twenty types of constructive and destructive control that were studied for approval voting [22] and two additional control types introduced by Faliszewski et al. [18], and we do so for a voting system that was proposed by Brams and Sanver [8] as a combination of preference-based and approval voting. Approval voting was introduced by Brams and Fishburn ([4, 5], see also [6]) as follows: Every voter either approves or disapproves of each candidate, and every candidate with the largest number of approvals is a winner. One of the simplest preference-based voting systems is plurality: All voters report their preference rankings of the candidates, and the winners are the candidates that are ranked first-place by the largest number of voters. The purpose of this paper is to show that Brams and Sanver's combined system (adapted here so as to keep its useful features even in the presence of control actions) combines the strengths, in terms of computational resistance to control, of plurality and approval voting.

Some voting systems are *immune* to certain types of control in the sense that it is never possible for the chair to reach his or her goal via the corresponding control action. Of course, immunity to any type of control is most desirable, as it unconditionally shields the voting system against this particular control type. Unfortunately, like most voting systems approval voting is *susceptible* (i.e., not immune) to many types of control, and plurality voting is susceptible to all types of control. However, and this was Bartholdi, Tovey, and Trick's brilliant insight [3], even for systems susceptible to control, the chair's task of controlling a given election may be too hard computationally (namely, NP-hard) for him or her to succeed. The voting system is then said to be *resistant* to this control type. If a voting system is susceptible to some type of control, but the chair's task can be solved in polynomial time, the system is said to be *vulnerable* to this control type.

The quest for a natural voting system with an easy winner-determination procedure that is universally resistant to control has lasted for more than 15 years now. Among the voting systems that have been studied with respect to control are plurality, Condorcet, approval, cumulative, Llull, and (variants of) Copeland voting [3, 22, 23, 27, 18, 19]. Among these systems, plurality and Copeland voting (denoted Copeland$^{0.5}$ in [19]) display the broadest resistance to control, yet even they are not universally control-resistant. The only system currently known to be fully resistant—to the 20 types of constructive and destructive control studied in [22, 23]—is a highly artificial system constructed via hybridization [23]. (We mention that this system was not designed for direct, real-world use as a "natural" system but rather was intended to rule out the existence of a certain impossibility theorem [23].)

While approval voting nicely distinguishes between each voter's acceptable and inacceptable candidates, it ignores the preference rankings the voters may have about their approved (or disapproved) candidates. This shortcoming motivated Brams and Sanver [8] to introduce a voting system that combines approval and preference-based voting, and they defined the related notions of sincere and admissible approval strategies, which are quite natural requirements. We adapt their sincere-strategy preference-based approval voting system in a natural way such that, for elections with at least two candidates, admissibility of approval strategies (see Definition 2.1) can be ensured even in the presence of control actions such as deleting candidates and partitioning candidates or voters. Note that, in control by partition of voters (see [14]), the run-off may have a reduced number of candidates.

The purpose of this paper is to study if, and to what extent, this hybrid system (where "hybrid" is not meant in the sense of [23] but refers to combining preference-based with approval voting in the sense of Brams and Sanver [8]) inherits the control resistances of plurality (which is perhaps the simplest preference-based system) and approval voting. Denoting this system by SP-AV, we show that SP-AV does combine all the resistances of plurality and approval voting.

More specifically, we prove that sincere-strategy preference-based approval voting is resistant to 19 and vulnerable to only three of the 22 types of control considered here. For

Number of	Condorcet	Approval	Llull	Copeland	Plurality	SP-AV
resistances	3	4	14	15	16	19
immunities	4	9	0	0	0	0
vulnerabilities	7	9	8	7	6	3
References	[3, 22]	[3, 22]	[18, 19]	[18, 19]	[3, 22, 18]	[13, 22] and this paper

Table 1: Number of resistances, immunities, and vulnerabilities to our 22 control types.

comparison, Table 1 shows the number of resistances, immunities, and vulnerabilities to our 22 control types that are known for each of Condorcet,[2] approval, Llull, plurality, and Copeland voting (see [3, 22, 18, 19]), and for SP-AV (see Theorem 3.1 and Table 2).

This paper is organized as follows. In Section 2, we define sincere-strategy preference-based approval voting, the types of control studied in this paper, and the notions of immunity, susceptibility, vulnerability, and resistance [3]. In Section 3, we prove our results on SP-AV. Finally, in Section 4 we give our conclusions and state some open problems.

2 Preliminaries

2.1 Preference-Based Approval Voting

An election $E = (C, V)$ is specified by a finite set C of candidates and a finite collection V of voters who express their preferences over the candidates in C, where distinct voters may of course have the same preferences. How the voter preferences are represented depends on the voting system used. In approval voting (AV, for short), every voter draws a line between his or her acceptable and inacceptable candidates (by specifying a 0-1 approval vector, where 0 represents disapproval and 1 represents approval), yet does not rank them. In contrast, many other important voting systems (e.g., Condorcet voting, Copeland voting, all scoring protocols including plurality, Borda count, veto, etc.) are based on voter preferences that are specified as tie-free linear orderings of the candidates. As is most common in the literature, votes will here be represented nonsuccinctly: one ballot per voter. Note that some papers (e.g., [16, 18, 19]) also consider succinct input representations for elections where multiplicities of votes are given in binary.

Brams and Sanver [8] introduced a voting system that combines approval and preference-based voting. To distinguish this system from other systems that these authors introduced with the same purpose of combining approval and preference-based voting [7], we call the variant considered here (including the conventions and rules to be explained below) *sincere-strategy preference-based approval voting* (SP-AV, for short).

Definition 2.1 ([8]) *Let (C, V) be an election, where the voters both indicate approvals/disapprovals of the candidates and provide a tie-free linear ordering of all candidates. For each voter $v \in V$, an* AV strategy of *v is a subset $S_v \subseteq C$ such that v approves of all candidates in S_v and disapproves of all candidates in $C - S_v$. The list of AV strategies*

[2] As in [22], we consider two types of control by partition of candidates (namely, with and without runoff) and one type of control by partition of voters, and for each partition case we use the rules TE ("ties eliminate") and TP ("ties promote") for handling ties that may occur in the corresponding subelections (see [14]). However, since Condorcet winners are always unique when they exist, the distinction between TE and TP is not made for the partition cases within Condorcet voting. Note further that the two additional control types in Section 2.2.1 (namely, constructive and destructive control by adding a limited number of candidates [18]) have not been considered for Condorcet voting [3, 22]. That is why Table 1 lists only 14 instead of 22 types of control for Condorcet.

for all voters in V is called an AV strategy profile for (C, V). *(We sometimes also speak of V's* AV strategy profile for C.*) For each $c \in C$, let $score_{(C,V)}(c) = \|\{v \in V \mid c \in S_v\}\|$ denote the number of c's approvals. Every candidate c with the largest $score_{(C,V)}(c)$ is a winner of election (C, V).*

An AV strategy S_v of a voter $v \in V$ is said to be admissible *if S_v contains v's most preferred candidate and does not contain v's least preferred candidate. S_v is said to be* sincere *if for each $c \in C$, if v approves of c then v also approves of each candidate ranked higher than c (i.e., there are no gaps allowed in sincere approval strategies). An AV strategy profile for (C, V) is* admissible *(respectively,* sincere*) if the AV strategies of all voters in V are admissible (respectively, sincere).*

Admissibility and sincerity are quite natural requirements. In particular, requiring the voters to be sincere ensures that their preference rankings and their approvals/disapprovals are not contradictory. Note further that admissible AV strategies are not dominated in a game-theoretic sense [4], and that sincere strategies for at least two candidates are always admissible if voters are neither allowed to approve of everybody nor to disapprove of everybody (i.e., if we require voters v to have only AV strategies S_v with $\emptyset \neq S_v \neq C$), a convention adopted by Brams and Sanver [8] and also adopted here.[3] Henceforth, we will tacitly assume that only sincere AV strategy profiles are considered (which by the above convention, whenever there are at least two candidates,[4] necessarily are admissible), i.e., a vote with an insincere strategy will be considered void.

Preferences are represented by a left-to-right ranking (separated by a space) of the candidates (e.g., $a \;\; b \;\; c$), with the leftmost candidate being the most preferred one, and approval strategies are denoted by inserting a straight line into such a ranking, where all candidates left of this line are approved and all candidates right of this line are disapproved (e.g., "$a \;\mid\; b \;\; c$" means that a is approved, while both b and c are disapproved). In our constructions, we sometimes also insert a subset $B \subseteq C$ into such approval rankings, where we assume some arbitrary, fixed order of the candidates in B (e.g., "$a \;\mid\; B \;\; c$" means that a is approved, while all $b \in B$ and c are disapproved).

2.2 Control Problems for Preference-Based Approval Voting

The control problems considered here were introduced by Bartholdi, Tovey, and Trick [3] for constructive control and by Hemaspaandra, Hemaspaandra, and Rothe [22] for destructive control. In constructive control scenarios the chair's goal is to make a favorite candidate win, and in destructive control scenarios the chair's goal is to ensure that a despised candidate does not win. As is common, the chair is assumed to have complete knowledge of the voters' preference rankings and approval strategies (see [22] for a detailed discussion of this assumption), and as in most papers on electoral control (exceptions are, e.g., [27, 19]) we define the control problems in the unique-winner model.

To achieve his or her goal, the chair modifies the structure of a given election via adding/deleting/partitioning either candidates or voters. Such control actions—specifically,

[3]Brams and Sanver [8] actually preclude only the case $S_v = C$ for voters v. However, an AV strategy that disapproves of all candidates obviously is sincere, yet not admissible, which is why we also exclude the case of $S_v = \emptyset$.

[4]Note that an AV strategy is never admissible for less than two candidates. We mention in passing that a precursor of this paper [13] specifically required for single-candidate elections that each voter must approve of this candidate. In this version of the paper, we drop this requirement for two reasons. First, it in fact is not needed because the one candidate in a single-candidate election will always win—even with zero approvals (i.e., SP-AV is a "voiced" voting system). Second, it is very well comprehensible that a voter, when given just a single candidate (think, for example, of an "election" in the Eastern bloc before 1989), can get some satisfaction from denying this candidate his or her approval, even if he or she knows that this disapproval won't prevent the candidate from winning.

those with respect to control via deleting or partitioning candidates or via partitioning voters—may have an undesirable impact on the resulting election in that they might violate our conventions about admissible AV strategies. That is why we define the following rule that preserves (or re-enforces) our conventions under such control actions:

> Whenever during or after a control action it happens that we obtain an election (C, V) with $\|C\| \geq 2$ and for some voter $v \in V$ we have $S_v = \emptyset$ or $S_v = C$, then each such voter's AV strategy is changed to approve of his or her top candidate and to disapprove of his or her bottom candidate. This rule re-enforces $\emptyset \neq S_v \neq C$ for each $v \in V$.

We now formally define those of our control problems that are relevant for the proofs we give later; for the definition of the remaining control problems, see the full version [14]. Each problem is defined by stating the problem instance together with two questions, one for the constructive and one for the destructive case. These control problems are tailored to sincere-strategy preference-based approval voting by requiring every election occuring in these control problems (be it before, during, or after a control action—so, in particular, this also applies to the subelections in the partitioning cases) to have a sincere AV strategy profile and to satisfy the above conventions and rules. In particular, this means that when the number of candidates is reduced (due to deleting candidates or partitioning candidates or voters), approval lines may have to be moved in accordance with the above rules.

2.2.1 Control by Adding Candidates

In this control scenario, the chair seeks to reach his or her goal by adding to the election, which originally involves only "qualified" candidates, some new candidates who are chosen from a given pool of spoiler candidates. In their study of control for approval voting, Hemaspaandra et al. [22] considered only the case of adding an *unlimited* number of spoiler candidates (which is the original variant of this problem as defined by Bartholdi et al. [3]). We consider the same variant of this problem here to make our results comparable with those established in [22], but for completeness we in addition consider the case of adding a *limited* number of spoiler candidates, where the prespecified limit is part of the problem instance. This variant of this problem was introduced by Faliszewski et al. [18, 19] in analogy with the definitions of control by deleting candidates and of control by adding or deleting voters. They showed that, for the election system Copeland$^\alpha$ they investigate, the complexity of these two problems can drastically change depending on the parameter α, see [19].

Name: Control by Adding an Unlimited Number of Candidates.
Instance: An election $(C \cup D, V)$ and a designated candidate $c \in C$, where the set C of qualified candidates and the set D of spoiler candidates are disjoint.
Question (constructive): Is it possible to choose a subset $D' \subseteq D$ such that c is the unique winner of election $(C \cup D', V)$?
Question (destructive): Is it possible to choose a subset $D' \subseteq D$ such that c is not a unique winner of election $(C \cup D', V)$?

The problem Control by Adding a Limited Number of Candidates is defined analogously, with the only difference being that the chair seeks to reach his or her goal by adding at most k spoiler candidates, where k is part of the problem instance.

2.2.2 Control by Deleting Candidates

In this control scenario, the chair seeks to reach his or her goal by deleting (up to a given number of) candidates. Here it may happen that our conventions are violated by the control

action, but will be re-enforced by the above rules (namely, by moving the line between some voter's acceptable and inacceptable candidates to behind the top candidate or to before the bottom candidate whenever necessary).

Name: Control by Deleting Candidates.
Instance: An election (C, V), a designated candidate $c \in C$, and a nonnegative integer k.
Question (constructive): Is it possible to delete up to k candidates from C such that c is the unique winner of the resulting election?
Question (destructive): Is it possible to delete up to k candidates (other than c) from C such that c is not a unique winner of the resulting election?

2.3 Immunity, Susceptibility, Vulnerability, and Resistance

Definition 2.2 ([3]) *Let $\mathcal{E}$ be an election system and let Φ be some given type of control. $\mathcal{E}$ is said to be* immune to Φ-control *if (a) Φ is a constructive control type and it is never possible for the chair to turn a designated candidate from being not a unique winner into being the unique winner via exerting Φ-control, or (b) Φ is a destructive control type and it is never possible for the chair to turn a designated candidate from being the unique winner into being not a unique winner via exerting Φ-control. $\mathcal{E}$ is said to be* susceptible to Φ-control *if it is not immune to Φ-control. $\mathcal{E}$ is said to be* vulnerable to Φ-control *if $\mathcal{E}$ is susceptible to Φ-control and the control problem associated with Φ is solvable in polynomial time. $\mathcal{E}$ is said to be* resistant to Φ-control *if $\mathcal{E}$ is susceptible to Φ-control and the control problem associated with Φ is* NP*-hard.*

For example, approval voting is known to be immune to eight of the twelve types of candidate control considered in [22]. The proofs of these results crucially employ the links between immunity/susceptibility for various control types shown in [22] and the fact that approval voting satisfies the unique version of the Weak Axiom of Revealed Preference (denoted by Unique-WARP, see [22, 3]): If a candidate c is the unique winner in a set C of candidates, then c is the unique winner in every subset of C that includes c. In contrast with approval voting, SP-AV does not satisfy Unique-WARP, and we will see later in Section 3.2 that it indeed is susceptible to each type of control considered here.

Proposition 2.3 *Sincere-strategy preference-based approval voting does not satisfy Unique-WARP.*

3 Results for SP-AV

3.1 Overview

Theorem 3.1 below (see also Table 2) shows the complexity results regarding control of elections for SP-AV. As mentioned in the introduction, with 19 resistances and only three vulnerabilities, this system has more resistances and fewer vulnerabilities to control (for our 22 control types) than is currently known for any other natural voting system with a polynomial-time winner problem.

Theorem 3.1 *Sincere-strategy preference-based approval voting is resistant and vulnerable to our 22 types of control as shown in Table 2.*

Control by	SP-AV		AV	
	Constr.	Destr.	Constr.	Destr.
Adding an Unlimited Number of Candidates	R	R	I	V
Adding a Limited Number of Candidates	R	R	I	V
Deleting Candidates	R	R	V	I
Partition of Candidates	TE: R TP: R	TE: R TP: R	TE: V TP: I	TE: I TP: I
Run-off Partition of Candidates	TE: R TP: R	TE: R TP: R	TE: V TP: I	TE: I TP: I
Adding Voters	R	V	R	V
Deleting Voters	R	V	R	V
Partition of Voters	TE: R TP: R	TE: V TP: R	TE: R TP: R	TE: V TP: V

Table 2: Overview of results. Key: I means immune, R means resistant, V means vulnerable, TE means ties-eliminate, and TP means ties-promote. Results for SP-AV are new. Results for AV, stated here to allow comparison, are due to Hemaspaandra, Hemaspaandra, and Rothe [22]. (The results for control by adding a limited number of candidates for approval voting, though not stated explicitly in [22], follow immediately from the proofs of the corresponding results for the "unlimited" variant of the problem.)

3.2 Susceptibility

By definition, all resistance and vulnerability results in particular require susceptibility. The following two lemmas (the proofs of which can be found in the full version [14]) show that SP-AV is susceptible to the 22 types of control we consider.

Lemma 3.2 *SP-AV is susceptible to constructive and destructive control by adding candidates (in both the "limited" and the "unlimited" variant of the problem), by deleting candidates, and by partition of candidates (with or without run-off and for each in both tie-handling models, TE and TP).*

Lemma 3.3 *SP-AV is susceptible to constructive and destructive control by adding voters, by deleting voters, and by partition of voters in both tie-handling models, TE and TP.*

3.3 Candidate Control

Theorems 3.4, 3.5, and 3.6 below show that sincere-strategy preference-based approval voting is fully resistant to candidate control. This result should be contrasted with that of Hemaspaandra, Hemaspaandra, and Rothe [22], who proved immunity and vulnerability for all cases of candidate control within approval voting (see Table 2). In fact, SP-AV has the same resistances to candidate control as plurality, and we note that the construction presented in [22] to prove plurality resistant also works for SP-AV in all cases of candidate control mentioned in Theorem 3.4.

All resistance results in this section follow via a reduction from the NP-complete problem Hitting Set (see, e.g., Garey and Johnson [20]): Given a set $B = \{b_1, b_2, \ldots, b_m\}$, a collection $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$ of subsets $S_i \subseteq B$, and a positive integer $k \leq m$, does $\mathcal{S}$ have a hitting set of size at most k, i.e., is there a set $B' \subseteq B$ with $\|B'\| \leq k$ such that for each i, $S_i \cap B' \neq \emptyset$?

Some of our proofs use constructions and arguments for SP-AV that are straightforward modifications of the constructions and arguments of the corresponding results for approval

voting or plurality from [22], whereas some other of our results require new insights to make the proof work for SP-AV. For completeness, we also state the results for SP-AV that follow by straightforward modifications of known constructions for approval or plurality, attributing them to Hemaspaandra et al. [22] (such as Theorem 3.4 below). Proof sketches of these results (explicitly showing the modifications needed to make the proofs work for SP-AV) can be found in the full version [14]. Results that are not explicitly attributed to Hemaspaandra et al. [22] use novel constructions or arguments specific to SP-AV.

Theorem 3.4 ([22]) *SP-AV is resistant to all types of constructive and destructive candidate control considered here except for (a) constructive control by deleting candidates and (b) constructive and destructive control by adding a limited number of candidates.*

We will explain in more detail below why case (a) in Theorem 3.4 is missing, and we will show as Theorem 3.5 that resistance holds for this case (a). Case (b) is missing in Theorem 3.4 simply because this control type ("adding a limited number of candidates") has not been considered in [22], but we will establish resistance for this missing case (b) as Theorem 3.6 below, via modifying our reduction presented in the proof of Theorem 3.5.

As to the missing case (a) mentioned in Theorem 3.4 above: Why does the construction that works for plurality (see [22] and also the full version [14]) not work to show that SP-AV is resistant to constructive control by deleting candidates? Informally put, the reason is that candidate c is the only serious rival of candidate w in the election (C, V) defined in Construction 4.28 of [22] (see also Construction 3.5 in [14]), so by simply deleting c the chair could make w the unique SP-AV winner, regardless of whether $\mathcal{S}$ has a hitting set of size k. However, via a different construction, we can prove resistance also in this case.

Theorem 3.5 *SP-AV is resistant to constructive control by deleting candidates.*

Proof. Susceptibility holds by Lemma 3.2. To prove resistance, we provide a reduction from Hitting Set. Let $(B, \mathcal{S}, k)$ be a given instance of Hitting Set, where $B = \{b_1, b_2, \ldots, b_m\}$ is a set, $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$ is a collection of subsets $S_i \subseteq B$, and $k < m$ is a positive integer.[5]

Define the election (C, V), where $C = B \cup \{w\}$ is the candidate set and where V consists of the following $4n(k+1) + 4m - 2k + 3$ voters:

1. For each i, $1 \leq i \leq n$, there are $2(k+1)$ voters of the form: $S_i \mid (B - S_i)\ w$.
2. For each i, $1 \leq i \leq n$, there are $2(k+1)$ voters of the form: $(B - S_i)\ w \mid S_i$.
3. For each j, $1 \leq j \leq m$, there are two voters of the form: $b_j \mid w\ (B - \{b_j\})$.
4. There are $2(m-k)$ voters of the form: $B \mid w$.
5. There are three voters of the form: $w \mid B$.

Since for each $b_j \in B$, the difference

$$score_{(C,V)}(w) - score_{(C,V)}(b_j) = 2n(k+1) + 3 - (2n(k+1) + 2 + 2(m-k)) = 1 - 2(m-k)$$

is negative (due to $k < m$), w loses to each member of B and so does not win election (C, V).

We claim that $\mathcal{S}$ has a hitting set B' of size k if and only if w can be made the unique SP-AV winner by deleting at most $m - k$ candidates.

[5] Note that if $k = m$ then B is always a hitting set of size at most k (provided that $\mathcal{S}$ contains only nonempty sets—a requirement that doesn't affect the NP-completeness of the problem), and we thus may require that $k < m$.

From left to right: Suppose $\mathcal{S}$ has a hitting set B' of size k. Then, for each $b_j \in B'$,

$$\begin{aligned} & score_{(B' \cup \{w\},V)}(w) - score_{(B' \cup \{w\},V)}(b_j) \\ &= \quad 2n(k+1) + 2(m-k) + 3 - (2n(k+1) + 2 + 2(m-k)) \;=\; 1, \end{aligned}$$

since the approval line is moved for $2(m-k)$ voters of the third group, thus transferring their approvals from members of $B - B'$ to w. So w is the unique SP-AV winner of election $(B' \cup \{w\}, V)$. Since $B' \cup \{w\} = C - (B - B')$, it follows from $\|B\| = m$ and $\|B'\| = k$ that deleting $m-k$ candidates from C makes w the unique SP-AV winner.

From right to left: Let $D \subseteq B$ be any set such that $\|D\| \leq m-k$ and w is the unique SP-AV winner of election $(C-D, V)$. Let $B' = (C-D) - \{w\}$. Note that $B' \subseteq B$ and that we have the following scores in $(B' \cup \{w\}, V)$:

$$\begin{aligned} score_{(B' \cup \{w\},V)}(w) &= 2(n-\ell)(k+1) + 2(m - \|B'\|) + 3, \\ score_{(B' \cup \{w\},V)}(b_j) &\leq 2n(k+1) + 2(k+1)\ell + 2 + 2(m-k) \quad \text{for each } b_j \in B', \end{aligned}$$

where ℓ is the number of sets $S_i \in \mathcal{S}$ that are not hit by B', i.e., $B' \cap S_i = \emptyset$. Since w is the unique SP-AV winner of $(B' \cup \{w\}, V)$, w has more approvals than any candidate b_j in B':

$$\begin{aligned} & score_{(B' \cup \{w\},V)}(w) - score_{(B' \cup \{w\},V)}(b_j) \\ &\geq \quad 2(n-\ell)(k+1) + 2(m - \|B'\|) + 3 - 2n(k+1) - 2\ell(k+1) - 2 - 2(m-k) \\ &= \quad 1 + 2(k - \|B'\|) - 4\ell(k+1) \;>\; 0. \end{aligned}$$

Solving this inequality for ℓ, we obtain $0 \leq \ell < \frac{1+2(k-\|B'\|)}{4(k+1)} < \frac{4+4k}{4(k+1)} = 1$. Thus $\ell = 0$. It follows that $1 + 2(k - \|B'\|) > 0$, which implies $\|B'\| \leq k$. Thus, B' is a hitting set of size at most k. ❑

Now consider the missing case (b) in Theorem 3.4. Control by adding a limited number of candidates has not been considered in [22], but we now prove resistance in this case by modifying the construction of the previous proof.

Theorem 3.6 *SP-AV is resistant to constructive and destructive control by adding a limited number of candidates.*

Proof. Susceptibility holds by Lemma 3.2 in both the constructive and destructive case.

To prove resistance in the constructive case, we slightly modify the construction presented in the proof of Theorem 3.5 that provides a reduction from Hitting Set. Let $(B, \mathcal{S}, k)$ be a given instance of Hitting Set, where $B = \{b_1, b_2, \ldots, b_m\}$ is a set, $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$ is a collection of subsets $S_i \subseteq B$, and $k < m$ is a positive integer (see Footnote 5 for why $k < m$ may be assumed).

We define the election (C, V) as in the proof of Theorem 3.5, except that we introduce a new candidate, a, and insert a into the preference lists of the voters. So $C = B \cup \{a, w\}$ is now the candidate set and V consists now of the following $4n(k+1) + 4m - 2k + 3$ voters:

1. For each i, $1 \leq i \leq n$, there are $2(k+1)$ voters of the form: $S_i \;\mid\; (B - S_i) \;\; a \;\; w$.
2. For each i, $1 \leq i \leq n$, there are $2(k+1)$ voters of the form: $(B - S_i) \;\; a \;\; w \;\mid\; S_i$.
3. For each j, $1 \leq j \leq m$, there are two voters of the form: $b_j \;\mid\; w \;\; (B - \{b_j\}) \;\; a$.
4. There are $2(m-k)$ voters of the form: $B \;\mid\; a \;\; w$.
5. There are three voters of the form: $w \;\mid\; a \;\; B$.

Our reduction maps the Hitting Set instance $(B, \mathcal{S}, k)$ to the instance $((C' \cup B, V), w, k)$ of Constructive Control by Adding a Limited Number of Candidates, where $C' = \{a, w\}$ is the set of qualified candidates, B is the set of spoiler candidates, w is the distinguished candidate, and k is the limit on the number of candidates that may be added. Note that $score_{(C',V)}(w) = 2m + 3$ and $score_{(C',V)}(a) = 4n(k+1) + 2(m-k)$. So $score_{(C',V)}(a) > score_{(C',V)}(w)$, and thus a is the unique winner of the election (C', V). However, by an argument analogous to that given in the proof of Theorem 3.5, we can show that $\mathcal{S}$ has a hitting set of size k if and only if w can be made the unique SP-AV winner by adding at most k candidates from B.

To prove resistance in the destructive case, we modify the above construction as follows: We map our Hitting Set instance $(B, \mathcal{S}, k)$ as above to the instance $((C' \cup B, V), w, k)$ of Destructive Control by Adding a Limited Number of Candidates, where the election (C, V) has the same candidate set $C = B \cup \{a, w\}$ as above, but V now consists of the following $4n(k+1) + 4m - 2k + 2$ voters:

1. For each i, $1 \leq i \leq n$, there are $2(k+1)$ voters of the form: $S_i \mid w \ (B - S_i) \ a$.
2. For each i, $1 \leq i \leq n$, there are $2(k+1)$ voters of the form: $(B - S_i) \ w \ a \mid S_i$.
3. For each j, $1 \leq j \leq m$, there are two voters of the form: $b_j \mid w \ (B - \{b_j\}) \ a$.
4. There are $2(m-k)$ voters of the form: $B \mid a \ w$.
5. There are two voters of the form: $w \mid a \ B$.

That is, the destructive case differs from the constructive case as follows: a and w have switched their positions in the voters of the first two groups and there are only two instead of three voters of the form $w \mid a \ B$ in the fifth group. In particular, w is now the unique winner of election (C', V), where $C' = \{a, w\}$, since $score_{(C',V)}(w) = 4n(k+1) + 2m + 2$ and $score_{(C',V)}(a) = 2(m-k)$. Again, by an argument analogous to that given in the proof of Theorem 3.5, we can show that $\mathcal{S}$ has a hitting set of size k if and only if it can be ensured by adding at most k candidates from B that w is not the unique SP-AV winner of the resulting election. This completes the proof. ❑

3.4 Voter Control

Turning now to voter control, most of the proofs for SP-AV follow from modifications of the corresponding constructions for approval voting given in [22], except for destructive control by partition of voters in model TE. Let us first state the former results.

Theorem 3.7 ([22]) *SP-AV is resistant to constructive control by adding voters and by deleting voters, to constructive and destructive control by partition of voters in model TP, and to constructive control by partition of voters in model TE. SP-AV is vulnerable to destructive control by adding voters and by deleting voters.*

Now, we turn to destructive control by partition of voters in model TE. While our polynomial-time algorithm showing vulnerability for SP-AV in this case is based on the corresponding polynomial-time algorithm for approval voting in [22], it extends their algorithm in a nontrivial way. The proof of Theorem 3.8 can be found in the full version [14].

Theorem 3.8 *SP-AV is vulnerable to destructive control by partition of voters in model TE.*

4 Conclusions and Open Questions

We have shown that Brams and Sanver's sincere-strategy preference-based approval voting system [8] combines the resistances of approval and plurality voting to procedural control: SP-AV is resistant to 19 of the 22 previously studied types of control. On the one hand, like Copeland voting [19], SP-AV is fully resistant to constructive control, yet unlike Copeland it additionally is broadly resistant to destructive control. On the other hand, like plurality [3, 22], SP-AV is fully resistant to candidate control, yet unlike plurality it additionally is broadly resistant to voter control. Thus, for these 22 types of control, SP-AV has more resistances, by three, and fewer vulnerabilities to control than is currently known for any other natural voting system with a polynomial-time winner problem.

As a work in progress, we are currently expanding our study of SP-AV's behavior with respect to procedural control towards other areas of computational social choice. In addition, we propose as an interesting and extremely ambitious task for future work the study of SP-AV (and other voting systems as well) beyond the worst-case—as we have done here—and towards an appropriate typical-case complexity model; see, e.g., [25, 26, 9, 24, 12] for interesting results and discussion in this direction.

Acknowledgments: We thank the anonymous MFCS-2008 and COMSOC-2008 referees for their helpful comments on preliminary versions of this paper.

References

[1] J. Bartholdi III, C. Tovey, and M. Trick. The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6(3):227–241, 1989.

[2] J. Bartholdi III, C. Tovey, and M. Trick. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare*, 6(2):157–165, 1989.

[3] J. Bartholdi III, C. Tovey, and M. Trick. How hard is it to control an election? *Mathematical Comput. Modelling*, 16(8/9):27–40, 1992.

[4] S. Brams and P. Fishburn. Approval voting. *American Political Science Review*, 72(3):831–847, 1978.

[5] S. Brams and P. Fishburn. *Approval Voting*. Birkhäuser, Boston, 1983.

[6] S. Brams and P. Fishburn. Voting procedures. In K. Arrow, A. Sen, and K. Suzumura, editors, *Handbook of Social Choice and Welfare*, volume 1, pages 173–236. North-Holland, 2002.

[7] S. Brams and R. Sanver. Voting systems that combine approval and preference. In S. Brams, W. Gehrlein, and F. Roberts, editors, *The Mathematics of Preference, Choice, and Order: Essays in Honor of Peter C. Fishburn*. Springer. To appear.

[8] S. Brams and R. Sanver. Critical strategies under approval voting: Who gets ruled in and ruled out. *Electoral Studies*, 25(2):287–305, 2006.

[9] V. Conitzer and T. Sandholm. Nonexistence of voting rules that are usually hard to manipulate. In *Proc. AAAI'06*, pages 627–634. AAAI Press, July 2006.

[10] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proc. WWW'01*, pages 613–622. ACM Press, 2001.

[11] E. Ephrati and J. Rosenschein. Multi-agent planning as a dynamic search for social consensus. In *Proc. IJCAI'93*, pages 423–429, 1993.

[12] G. Erdélyi, L. Hemaspaandra, J. Rothe, and H. Spakowski. On approximating optimal weighted lobbying, and frequency of correctness versus average-case polynomial time. In *Proc. FCT'07*, pages 300–311. Springer-Verlag LNCS #4639, August 2007.

[13] G. Erdélyi, M. Nowak, and J. Rothe. Sincere-strategy preference-based approval voting broadly resists control. In *Proc. MFCS'08*. Springer-Verlag LNCS, August 2008. To appear.

[14] G. Erdélyi, M. Nowak, and J. Rothe. Sincere-strategy preference-based approval voting fully resists constructive control and broadly resists destructive control. Technical Report cs.GT/0806.0535, ACM Computing Research Repository (CoRR), June 2008.

[15] R. Fagin, R. Kumar, and D. Sivakumar. Efficient similarity search and classification via rank aggregation. In *Proc. ACM SIGMOD Intern. Conf. on Management of Data*, pages 301–312. ACM Press, 2003.

[16] P. Faliszewski, E. Hemaspaandra, and L. Hemaspaandra. The complexity of bribery in elections. In *Proc. AAAI'06*, pages 641–646. AAAI Press, 2006.

[17] P. Faliszewski, E. Hemaspaandra, L. Hemaspaandra, and J. Rothe. A richer understanding of the complexity of election systems. In S. Ravi and S. Shukla, editors, *Fundamental Problems in Computing: Essays in Honor of Professor Daniel J. Rosenkrantz*. Springer. To appear.

[18] P. Faliszewski, E. Hemaspaandra, L. Hemaspaandra, and J. Rothe. Llull and Copeland voting broadly resist bribery and control. In *Proc. AAAI'07*, pages 724–730. AAAI Press, 2007.

[19] P. Faliszewski, E. Hemaspaandra, L. Hemaspaandra, and J. Rothe. Copeland voting fully resists constructive control. In *Proc. AAIM'08*, pages 165–176. Springer-Verlag LNCS #5034, June 2008.

[20] M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, New York, 1979.

[21] S. Ghosh, M. Mundhe, K. Hernandez, and S. Sen. Voting for movies: The anatomy of recommender systems. In *Proc. 3rd Annual Conference on Autonomous Agents*, pages 434–435. ACM Press, 1999.

[22] E. Hemaspaandra, L. Hemaspaandra, and J. Rothe. Anyone but him: The complexity of precluding an alternative. *Artificial Intelligence*, 171(5–6):255–285, 2007.

[23] E. Hemaspaandra, L. Hemaspaandra, and J. Rothe. Hybrid elections broaden complexity-theoretic resistance to control. In *Proc. IJCAI'07*, pages 1308–1314. AAAI Press, 2007.

[24] C. Homan and L. Hemaspaandra. Guarantees for the success frequency of an algorithm for finding Dodgson-election winners. *Journal of Heuristics*. To appear. Full version available as Department of Computer Science, University of Rochester Technical Report TR-881, September 2005, revised June 2007.

[25] J. McCabe-Dansted, G. Pritchard, and A. Slinko. Approximability of Dodgson's rule. *Social Choice and Welfare*, 2008. DOI 10.1007/s00355-007-0282-8.

[26] A. Procaccia and J. Rosenschein. Junta distributions and the average-case complexity of manipulating elections. *Journal of Artificial Intelligence Research*, 28:157–181, 2007.

[27] A. Procaccia, J. Rosenschein, and A. Zohar. Multi-winner elections: Complexity of manipulation, control, and winner-determination. In *Proc. IJCAI'07*, pages 1476–1481. AAAI Press, 2007.

Gábor Erdélyi, Markus Nowak, and Jörg Rothe
Institut für Informatik
Heinrich-Heine-Universität Düsseldorf
40225 Düsseldorf, Germany

Llull and Copeland Voting Computationally Resist Bribery and Control[1]

Piotr Faliszewski, Edith Hemaspaandra, Lane A. Hemaspaandra, Jörg Rothe

Abstract

Control and bribery are settings in which an external agent seeks to influence the outcome of an election. Constructive control of elections refers to attempts by an agent to, via such actions as addition/deletion/partition of candidates or voters, ensure that a given candidate wins [6]. Destructive control refers to attempts by an agent to, via the same actions, preclude a given candidate's victory [26]. An election system in which an agent can affect the result and in which recognizing the inputs on which the agent can succeed is NP-hard (polynomial-time solvable) is said to be resistant (vulnerable) to the given type of control. Aside from election systems with an NP-hard winner problem, the only systems previously known to be resistant to all the standard control types are highly artificial election systems created by hybridization [27].

We study a parameterized version of Copeland voting, denoted by Copeland$^\alpha$, where the parameter α is a rational number between 0 and 1 that specifies how ties are valued in the pairwise comparisons of candidates. In every previously studied constructive or destructive control scenario, we determine which of resistance or vulnerability holds for Copeland$^\alpha$ for each rational α, $0 \leq \alpha \leq 1$. In particular, we prove that Copeland$^{0.5}$, the system commonly referred to as "Copeland voting," provides full resistance to constructive control, and we prove the same for Copeland$^\alpha$, for all rational α, $0 < \alpha < 1$. Among systems with a polynomial-time winner problem, Copeland voting is the first natural election system proven to have full resistance to constructive control. In addition, we prove that both Copeland^0 and Copeland^1 (interestingly, Copeland^1 is an election system developed by the thirteenth-century mystic Ramon Llull) are resistant to all standard types of constructive control other than one variant of addition of candidates. Moreover, we show that for each rational α, $0 \leq \alpha \leq 1$, Copeland$^\alpha$ voting is fully resistant to bribery attacks, and we establish fixed-parameter tractability of bounded-case control for Copeland$^\alpha$.

We also study Copeland$^\alpha$ elections under more flexible models such as microbribery and extended control, we integrate the potential irrationality of voter preferences into many of our results, and we prove our results in both the unique-winner and the nonunique-winner model. Our vulnerability results for microbribery are proven via a novel technique involving min-cost network flow.

1 Introduction

Elections have played an important role in human societies for thousands of years. For example, elections were of central importance in the democracy of ancient Athens. There citizens typically could only agree (vote *yes*) or disagree (vote *no*) with the speaker, and simple majority-rule was in effect. The mathematical study of elections, give or take a

[1] Supported in part by DFG grants RO-1202/{9-3, 11-1, 12-1}, NSF grants CCR-0311021, CCF-0426761, and IIS-0713061, the Alexander von Humboldt Foundation's TransCoop program, the European Science Foundation's EUROCORES program LogICCC, and two Friedrich Wilhelm Bessel Research Awards. Work done in part while the first three authors were visiting Heinrich-Heine-Universität Düsseldorf and while the fourth author was visiting the University of Rochester. Some results have been presented at the 22nd AAAI Conference on Artificial Intelligence (AAAI-07) [20], and at the 4th International Conference on Algorithmic Aspects in Information and Management (AAIM-08) [21].

few discussions by the ancient Greeks and Romans, was until recently thought to have been initiated only a few centuries ago, namely in the breakthrough work of Borda and Condorcet—later in part reinvented by Dodgson (see, e.g., [31] for reprints of their classic papers). One of the most interesting results of this early work is Condorcet's observation [7] that if one conducts elections with more than two alternatives then even if all voters have rational (i.e., transitive) preferences, the society in aggregate can be irrational (indeed, can have cycles of strict preference). Based on his observations, Condorcet suggested that if there exists a candidate c such that c defeats any other candidate in a head-to-head contest then that candidate should win the election. Such a candidate is called a Condorcet winner. Clearly, there can be at most one Condorcet winner in any election and there might be none.

This understanding of history has been reconsidered during the past few decades, as it has been rediscovered that the study of elections was in fact considered deeply as early as the thirteenth century (see Hägele and Pukelsheim [24] and the citations therein regarding Ramon Llull and the fifteenth-century figure Cusanus, especially the citations that there are numbered 3, 5, and 24–27). Ramon Llull (b. 1232, d. 1315), a Catalan mystic, missionary, and philosopher developed an election system that (a) has an efficient winner-determination procedure and (b) elects a Condorcet winner whenever one exists and otherwise elects candidates that are, in some sense, closest to being Condorcet winners. Llull's motivation for developing an election system was to obtain a method of choosing the abbesses, abbots, bishops, and perhaps even the pope. His election ideas never gained public acceptance in medieval Europe and were long forgotten.

It is interesting to note that Llull allowed voters to have *irrational* preferences. Given three candidates, c, d, and e, it was perfectly acceptable for a voter to prefer c to d, d to e, and e to c. On the other hand, in modern studies of voting and election systems each voter's preferences are most typically modeled as a linear order over all candidates. (In this paper, as is standard, "linear order" implies strictness, i.e., no tie in the ordering.) Yet irrationality is a very tempting and natural concept.

Llull's election system is remarkably similar to what is now known as "Copeland elections" [11], a more than half-century old voting procedure that is based on pairwise comparisons of candidates: The winner (by a majority of votes—in this paper "majority" always, as is standard, means strict majority) of each such a head-to-head contest is awarded one point and the loser receives no point; in ties, both parties are (in the most common interpretation of Copeland's meaning) awarded half a point; whoever collects the most points over all these contests (including tie-related points) is the election's winner. In fact, the points awarded for ties in such head-to-head majority-rule contests are treated in two ways, half a point (most common) and zero points (less common), in the literature when speaking of Copeland elections. To provide a framework that can capture both those notions, as well as Llull's system and the whole family of systems created by choices of how we value ties, we introduce a parameterized version of Copeland elections in Definition 1.1 below.

An election is specified by a finite set C of candidates and a finite collection V of voters, where each voter has preferences over the candidates. We consider both rational and irrational voters. The preferences of a rational voter are expressed by a preference list of the form $a > b > c$ (assuming $C = \{a, b, c\}$), where the underlying relation $>$ is a transitive linear order. The preferences of an irrational voter are expressed by a preference table that for any two distinct candidates specifies which of them is preferred to the other by this voter. An election system is a rule that determines the winner(s) of each given election (C, V).

Definition 1.1 *Let α, $0 \leq \alpha \leq 1$, be a fixed rational number. In a* Copeland$^\alpha$ *election the voters indicate which among any two distinct candidates they prefer. For each such head-to-head contest, if some candidate is preferred by a majority of voters then he or she obtains one point and the other candidate obtains zero points, and if a tie occurs then both*

candidates obtain α points. Let $E = (C, V)$ be an election. For each $c \in C$, $score_E^\alpha(c)$ is the sum of c's Copeland$^\alpha$ *points in E. Every candidate c with maximum $score_E^\alpha(c)$ wins.*

Copeland$^\alpha_{\text{Irrational}}$ *denotes the same system but with voters allowed to be irrational.*

So the system widely referred to in the literature as "Copeland elections" is Copeland$^{0.5}$, where tied candidates receive half a point each (see, e.g., Merlin and Saari [36, 32]; the definition used by Conitzer et al. [10] can be scaled to be equivalent to Copeland$^{0.5}$). Copeland^0, where tied candidates come away empty-handed, has sometimes also been referred to as "Copeland elections" (see, e.g., Procaccia, Rosenschein, and Kaminka [34] and an early version of this paper [20]). The above-mentioned election system by Ramon Llull is in this notation nothing other than Copeland^1, where tied candidates are awarded one point each, just like winners of head-to-head contests.[2] The group stage of FIFA World Cup finals is in essence a collection of Copeland$^\alpha$ tournaments with $\alpha = 1/3$.

At first glance, one might be tempted to think that the definitional perturbation due to the parameter α in Copeland$^\alpha$ elections is negligible. However, it in fact can make the dynamics of Llull's system quite different from those of, for instance, Copeland$^{0.5}$ or Copeland^0. We also mention that a probabilistic variant of Copeland voting was defined already in 1929 by Zermelo [38] and later on was reintroduced by several other researchers.

In general it is impossible to design a perfect election system. In the 1950s Arrow [2] famously showed that there is no social choice system that satisfies a certain small set of reasonable requirements, and later Gibbard [23], Satterthwaite [37], and Duggan and Schwartz [13] showed that any natural election system can be manipulated by strategic voting, i.e., by a voter who reveals different preferences than his or her true ones in order to affect an election's result in his or her favor. Also, no natural election system with a polynomial-time winner-determination procedure has yet been shown to be resistant to all types of control via procedural changes. Control refers to attempts by an external agent (called "the chair") to, via such actions as addition/deletion/partition of candidates or voters, make a given candidate win the election (in the case of constructive control [6]) or preclude a given candidate's victory (in the case of destructive control [26]).

These obstacles are very discouraging, but the field of computational social choice theory grew in part from the realization that computational complexity provides a tool to partially circumvent these obstacles. In particular, around 1990 Bartholdi, Tovey, and Trick [4, 6] and Bartholdi and Orlin [3] brilliantly observed that while we might not be able to make manipulation (i.e., strategic voting) and control of elections impossible, we can at least try to make such manipulation and control so computationally difficult that neither voters nor election organizers will attempt it. For example, if there is a way for a committee's chair to set up an election within the committee in such a way that his or her favorite option is guaranteed to win but the chair's computational task would take a million years, then for all practical purposes we may assume that the chair is prevented from finding such a set-up.

Since the seminal work of Bartholdi, Orlin, Tovey, and Trick a large body of research has been dedicated to the study of computational properties of election systems. Some topics that have received much attention are the complexity of manipulating elections [8, 9, 10, 14, 25, 33, 35] and of controlling elections via procedural changes [26, 27, 35, 17]. Recently,

[2] Page 23 of Hägele and Pukelsheim [24] indicates in a way we find deeply convincing (namely by a direct quote of Llull's in-this-case-very-clear words from his *Artifitium Electionis Personarum*—which was rediscovered by those authors in the year 2000) that at least one of Llull's election systems was Copeland^1, and so in this paper we refer to the both-candidates-score-a-point-on-a-tie variant as Llull voting.

In some settings Llull required the candidate and voter sets to be identical and had an elaborate two-stage tie-breaking rule ending in randomization. We disregard these issues here and cast his system into the modern idiom for election systems. (However, we note in passing that there do exist some modern papers in which the voter and candidate sets are taken to be identical, see for example the work of and references in [1].)

Faliszewski, Hemaspaandra, and Hemaspaandra introduced the study of the complexity of bribery in elections ([19], see also [18]). Bribery shares some features of manipulation and some features of control. In particular, the briber picks the voters he or she wants to affect (as in voter control problems) and asks them to vote as he or she wishes (as in manipulation).

The goal of this paper is to study Copeland$^\alpha$ elections from the point of view of computational social choice theory, in the setting where voters are rational and in the setting where the voters are allowed to have irrational preferences. (Note: When we henceforth say "irrational voters," we mean that the voters may have irrational preferences, not that they each must.) We study the issues of bribery and control and we point the reader to the work of Faliszewski, Hemaspaandra, and Schnoor [22] for work on manipulation.

A standard technique for showing that a particular election-related problem (e.g., the problem of deciding whether the chair can make his or her favorite candidate a winner by influencing at most k voters not to cast their votes) is computationally intractable is to show that it is NP-hard. This approach is taken in almost all of the papers on computational social choice cited above, and it is the approach that we take in this paper. One of the justifications for using NP-hardness as a barrier against manipulation and control of elections is that in multiagent settings any attempts to influence the election's outcome are made by computationally bounded software agents that have neither human intuition nor the computational ability to solve NP-hard problems.

Recently, such papers as [30, 33, 9, 28] have studied the frequency (or sometimes, probability weight) of correctness of heuristics for voting problems. We view worst-case study as a natural prerequisite to a frequency-of-hardness attack: After all, there is no point in seeking frequency-of-hardness results if the problem at hand is in P to begin with. And if one cannot even prove worst-case hardness for a problem, then proving average-case hardness is even more beyond reach. Also, current frequency results have debilitating limitations (for example, being locked into specific distributions; depending on unproven assumptions; adopting "tractability" notions that declare undecidable problems tractable and that are not robust under even linear-time reductions). Although frequency of hardness is a fascinating and important direction, these models are arguably not ready for prime time and, contrary to some people's impression, fail to imply average-case polynomial runtime claims. [15, 28] provide discussion of some of these issues.

We also mention that during our study of Copeland control we have noticed that the proof of an important result of Bartholdi, Tovey, and Trick [6, Theorem 12] (namely, that Condorcet voting is resistant to constructive control by deleting voters) is invalid. The invalidity is due to the proof centrally using nonstrict voters, in violation of Bartholdi, Tovey, and Trick's [6] (and our) model, and the invalidity seems potentially daunting to seek to fix with the proof approach taken there. We noticed also that Theorem 14 of the same paper has a similar flaw, and we have validly reproven their claimed results using our techniques (see [20] and the in-preparation full version of this paper).

Due to space limitations all proofs in this paper are omitted.

2 Bribery

In this section we present our results on the complexity of bribery for the Copeland$^\alpha$ election systems, where α is a rational number with $0 \leq \alpha \leq 1$. Our main result, which will be presented in Section 2.1, is that each such system is resistant to bribery, regardless of voters' rationality and of our mode of operation (constructive versus destructive). In Section 2.2, we will provide vulnerability results for Llull and Copeland^0 with respect to "microbribery."

2.1 Resistance to Bribery

Let $\mathcal{E}$ be an election system. In our case, $\mathcal{E}$ will be either Copeland$^\alpha$ or Copeland$^\alpha_{\text{Irrational}}$, where α, $0 \leq \alpha \leq 1$, is a fixed rational number. The bribery problem for $\mathcal{E}$ with rational voters is defined as follows [19].

Name: $\mathcal{E}$-bribery and $\mathcal{E}$-destructive-bribery.

Given: A set C of candidates, a collection V of voters specified via their preference lists over C, a distinguished candidate $p \in C$, and a nonnegative integer k.

Question (constructive): Is it possible to make p a winner of the $\mathcal{E}$ election resulting from (C, V) by modifying the preference lists of at most k voters?

Question (destructive): Is it possible to ensure that p is not a winner of the $\mathcal{E}$ election resulting from (C, V) by modifying the preference lists of at most k voters?

Our bribery problems are defined above for rational voters only and in the *nonunique-winner* model, i.e., asking whether a given candidate can be made, or prevented from being, *a* winner. Nonetheless, we have proven all our bribery results (and all our control results as well) both for the case of *nonunique winners* and *unique winners*. In the unique-winner model, we ask whether a given candidate can be made, or prevented from being, the sole winner. The versions of these problems for elections with irrational voters allowed is defined exactly as the rational one, with the only difference being that voters are represented via preference tables rather than preference lists, and the briber may completely change a voter's preference table at unit cost.

Theorem 2.1 *For each rational α, $0 \leq \alpha \leq 1$,* Copeland$^\alpha$ *is resistant to both constructive and destructive bribery in both the rational-voters case and the irrational-voters case, in both the nonunique-winner model and in the unique-winner model.*

2.2 Vulnerability to Microbribery for Irrational Voters

In this section we explore the problems related to microbribery of irrational voters. In standard bribery problems, which were considered in Section 2.1, we ask whether it is possible to ensure that a designated candidate p is a winner (or, in the destructive case, to ensure that p is not a winner) via modifying the preference tables of at most k voters. That is, we can at unit cost completely redefine the preference table of each voter bribed. Often such an approach is right: We pay for a service (namely, the modification of the reported preference table) and we pay for it in bulk (when we buy a voter, we have complete control over his or her preferences). However, sometimes it may be more reasonable to adopt a more local approach and to pay separately for each preference-table entry flip.

For each rational α, $0 \leq \alpha \leq 1$, we define the following two problems.

Name: Copeland$^\alpha_{\text{Irrational}}$-microbribery and Copeland$^\alpha_{\text{Irrational}}$-destructive-microbribery.

Given: A set C of candidates, a collection V of voters specified via their preference tables over C, a distinguished candidate $p \in C$, and a nonnegative integer k.

Question (constructive): Is it possible to make p a winner of the election resulting from (C, V) by flipping at most k entries in the preference tables of voters in V?

Question (destructive): Is it possible to guarantee that p is not a winner of the election resulting from (C, V) by flipping at most k entries in the preference tables of voters in V?

Our first result regarding microbribery is that destructive microbribery is easy for $\text{Copeland}^{\alpha}_{\text{Irrational}}$, for each rational α, $0 \leq \alpha \leq 1$. We take this opportunity to remind the reader that although the definition of vulnerability requires only that there be a polynomial-time algorithm to determine whether a successful action (in the present case, a destructive microbribe) *exists*, we will in each vulnerability proof provide something far stronger, namely a polynomial-time algorithm that both determines whether a successful action exists and that, when so, explicitly finds a successful action.

Theorem 2.2 *For each rational α, $0 \leq \alpha \leq 1$, $\text{Copeland}^{\alpha}_{\text{Irrational}}$ is vulnerable to destructive microbribery.*

Theorem 2.2 is proven via greedy algorithms. The constructive case is more complicated, but we still are able to obtain, for the values $\alpha \in \{0, 1\}$, polynomial-time algorithms via a fairly involved use of flow networks to model how particular Copeland^{α} points travel between candidates.

Theorem 2.3 *For $\alpha \in \{0, 1\}$, $\text{Copeland}^{\alpha}_{\text{Irrational}}$ is vulnerable to constructive microbribery.*

3 Control

3.1 Example of one Control Problem and Our Naming Scheme

We now give an example of how to define the control problems we consider, in both the constructive and the destructive version. Let $\mathcal{E}$ be an election system. In our case, $\mathcal{E}$ will be either Copeland^{α} or $\text{Copeland}^{\alpha}_{\text{Irrational}}$, where α, $0 \leq \alpha \leq 1$, is a fixed rational number. The types of control we consider here are well-known from the literature (see, e.g., [6, 20, 26]) and we will content ourselves with the definition of only control via adding candidates. Note that there are two versions of this control type. The *unlimited* version (which, for the constructive case, was introduced by Bartholdi, Tovey, and Trick [6]) asks whether the election chair can add (any number of) candidates from a given pool of spoiler candidates in order to either make his or her favorite candidate win the election (in the constructive case), or prevent his or her despised candidate from winning (in the destructive case):

Name: $\mathcal{E}$-CCAC_{u} and $\mathcal{E}$-DCAC_{u}.

Given: Disjoint candidate sets C and D, a collection V of voters represented via their preference lists (or preference tables in the irrational case) over the candidates in $C \cup D$, and a distinguished candidate $p \in C$.

Constructive Question ($\mathcal{E}$-CCAC_{u}): Does there exist a subset D' of D such that p is a winner of the $\mathcal{E}$ election with candidates $C \cup D'$ and voters V?

Destructive Question ($\mathcal{E}$-DCAC_{u}): Does there exist a subset D' of D such that p isn't a winner of the $\mathcal{E}$ election with candidates $C \cup D'$ and voters V?

The only difference in the *limited* version of constructive and destructive control via adding candidates ($\mathcal{E}$-CCAC and $\mathcal{E}$-DCAC, for short) is that the chair needs to achieve his or her goal by adding at most k candidates from the given set of spoiler candidates. This version of control by adding candidates was proposed in [20] to synchronize the definition of control by adding candidates with the definitions of control by deleting candidates, adding voters, and deleting voters.

As seen in the above definition example, we use the following naming conventions for control problems. The name of a control problem starts with the election system used (when

clear from context, it may be dropped), followed by CC for "constructive control" or by DC for "destructive control," followed by the acronym of the type of control: AC for "adding (a limited number of) candidates," AC_u for "adding (an unlimited number of) candidates," DC for "deleting candidates," PC for "partition of candidates," RPC for "run-off partition of candidates," AV for "adding voters," DV for "deleting voters," and PV for "partition of voters," and all the partitioning cases (PC, RPC, and PV) are followed by the acronym of the tie-handling rule used in subelections, namely TP for "ties promote" (i.e., all winners of a given subelection are promoted to the final round of the election) and TE for "ties eliminate" (i.e., if there is more than one winner in a given subelection then none of this subelection's winners is promoted to the final round of the election).

Note that our definitions focus on *a winner*, i.e., they are in the *nonunique-winner model*. The *unique-winner* analogs of these problems can be defined by requiring the distinguished candidate p to be the unique winner (or to not be a unique winner in the destructive case).

Let $\mathcal{E}$ be an election system and let Φ be a control type. We say $\mathcal{E}$ is *immune to* Φ*-control* if the chair can never reach his or her goal (of making a given candidate win in the constructive case, and of blocking a given candidate from winning in the destructive case) via asserting Φ-control. $\mathcal{E}$ is said to be *susceptible to* Φ*-control* if $\mathcal{E}$ is not immune to Φ-control. $\mathcal{E}$ is said to be *vulnerable to* Φ*-control* if it is susceptible to Φ-control and there is a polynomial-time algorithm for solving the control problem associated with Φ. $\mathcal{E}$ is said to be *resistant to* Φ*-control* if it is susceptible to Φ-control and the control problem associated with Φ is NP-hard. The above notions were introduced by Bartholdi, Tovey, and Trick [6] (see also, e.g., [26, 35, 27, 20]).

3.2 Resistance and Vulnerability to Control

Our main result in this section is Theorem 3.1 below.

Theorem 3.1 *Let α be a rational number with $0 \leq \alpha \leq 1$.* Copeland$^\alpha$ *elections are resistant and vulnerable to control types as indicated in Table 1. The same results hold for the case of irrational voters and in both the nonunique-winner model and the unique-winner model.*

In particular, Theorem 3.1 says that the notion widely referred to in the literature simply as "Copeland elections," which we here for clarity call Copeland$^{0.5}$, possesses all ten of our basic types of constructive resistance and, in addition, even has constructive AC_u resistance. And the other notion that in the literature is occasionally referred to as "Copeland elections," namely Copeland^0, as well as Llull elections, which are here denoted by Copeland^1, both possess all ten of our basic types of constructive resistance. However, Copeland^0 and Copeland^1 are vulnerable to this eleventh type of constructive control, the incongruous but historically resonant notion of constructive control by adding an unlimited number of candidates (i.e., $CCAC_u$).

Note that Copeland$^{0.5}$ has a higher number of constructive resistances, by three, than even plurality, which was before this paper the reigning champ among natural election systems. (Although the results regarding plurality in Table 1 are stated for the unique-winner version of control, for all the table's Copeland$^\alpha$ cases, $0 \leq \alpha \leq 1$, our results hold both in the cases of unique winners and of nonunique winners, thus allowing an apples-to-apples comparison to hold.) Admittedly, plurality does perform better with respect to destructive candidate control problems, but still our study of Copeland$^\alpha$ makes significant steps forward in the quest for a fully control-resistant natural election system with an easy winner problem.

Among the systems with a polynomial-time winner problem, Copeland$^{0.5}$—and indeed all Copeland$^\alpha$, $0 < \alpha < 1$—have the most resistances currently known for any natural election system whose voters vote by giving preference lists. However, we mention that

	Copeland$^\alpha$						Plurality	
	$\alpha = 0$		$0 < \alpha < 1$		$\alpha = 1$			
Control type	CC	DC	CC	DC	CC	DC	CC	DC
AC_u	V	V	R	V	V	V	R	R
AC	R	V	R	V	R	V	R	R
DC	R	V	R	V	R	V	R	R
RPC-TP	R	V	R	V	R	V	R	R
RPC-TE	R	V	R	V	R	V	R	R
PC-TP	R	V	R	V	R	V	R	R
PC-TE	R	V	R	V	R	V	R	R
PV-TE	R	R	R	R	R	R	V	V
PV-TP	R	R	R	R	R	R	R	R
AV	R	R	R	R	R	R	V	V
DV	R	R	R	R	R	R	V	V

Table 1: Comparison of control results for Copeland$^\alpha$ elections, where α with $0 \leq \alpha \leq 1$ is a rational number, and for plurality-rule elections. R means resistance to a particular control type and V means vulnerability. The results regarding plurality are due to Bartholdi, Tovey, and Trick [6] and Hemaspaandra, Hemaspaandra, and Rothe [26]. (Note that CCAC and CCDC resistance results for plurality, not handled explicitly in [6, 26], follow immediately from the respective $CCAC_u$ and $DCAC_u$ results.)

after our work, Erdélyi, Nowak, and Rothe [17] have shown that a certain rather subtle election system with a richer voter preference type—each voter specifies both a permutation and a set—has nineteen (out of a possible twenty-two) control resistances.

3.3 FPT Algorithm Schemes for Bounded-Case Control

The study of fixed-parameter complexity (see, e.g., [12]) has been expanding explosively since it was parented as a field by Downey, Fellows, and others in the late 1980s and the 1990s. Although the area has built a rich variety of complexity classes regarding parameterized problems, for the purpose of the current paper we need focus only on one very important class, namely, the class FPT. Briefly put, a problem parameterized by some value j is said to be *fixed-parameter tractable* (equivalently, to belong to the class FPT) if there is an algorithm for the problem whose running time is $f(j)n^{O(1)}$.

Fixed–Parameter Tractability Results. In their seminal paper on NP-hard winner-determination problems, Bartholdi, Tovey, and Trick [5] suggested considering hard election problems for the cases of a bounded number of candidates or a bounded number of voters, and they obtained efficient-algorithm results for such cases. Within the study of elections, this same approach—seeking efficient fixed-parameter algorithms—has also been used, for example, within the study of bribery [19].

We obtain for resistant-in-general control problems a broad range of efficient algorithms for the case when the number of candidates or voters is bounded. Our algorithms are not merely polynomial time. Rather, we give algorithms that prove membership in FPT. And we prove that our FPT claims hold even under the succinct input model—in which the voters are input via "(preference-list, binary-integer-giving-frequency-of-that-preference-list)" pairs— and even in the case of irrational voters. We obtain such algorithms for all the voter-control cases, both for bounded numbers of candidates and for bounded numbers of voters, and for all the candidate-control cases with bounded numbers of candidates. On

the other hand, we show that for the resistant-in-general irrational-voter, candidate-control cases, resistance still holds even if the number of voters is limited to being at most two.

Let us fix our notation. We consider two parameterizations: bounding the number of candidates and bounding the number of voters. We use the same notations used throughout this paper to describe problems, except we postpend a "-BV$_j$" to a problem name to state that the number of voters may be at most j, and we postpend a "-BC$_j$" to a problem name to state that the number of candidates may be at most j. In each case, the bound applies to the full number of such items involved in the problem. For example, in the case of control by adding voters, the j must bound the total of the number of voters in the election added together with the number of voters in the pool of voters available for adding.

To state our fixed-parameter tractability results concisely, we borrow a notational approach from transformational grammar, and use square brackets as an "independent choice" notation. So, for example, the claim $\begin{bmatrix}\text{It}\\ \text{She}\\ \text{He}\end{bmatrix}\ \begin{bmatrix}\text{runs}\\ \text{walks}\end{bmatrix}$ is a shorthand for six assertions: It runs; She runs; He runs; It walks; She walks; and He walks. A special case is the symbol "$\emptyset$" which, when it appears in such a bracket, means that when unwound it should be viewed as no text at all. For example, "$\begin{bmatrix}\text{Succinct}\\ \emptyset\end{bmatrix}$ Copeland is fun" asserts both "Succinct Copeland is fun" and "Copeland is fun."

Theorem 3.2 *For each rational α, $0 \leq \alpha \leq 1$, and each choice from the independent choice brackets below, the specified parameterized (as j varies over $\mathbb{N}$) problem is in* FPT*:*

$$\begin{bmatrix}\text{succinct}\\ \emptyset\end{bmatrix}\text{-}\begin{bmatrix}\text{Copeland}^{\alpha}\\ \text{Copeland}^{\alpha}_{\text{Irrational}}\end{bmatrix}\text{-}\begin{bmatrix}\text{C}\\ \text{D}\end{bmatrix}\text{C}\begin{bmatrix}\text{AV}\\ \text{DV}\\ \text{PV-TE}\\ \text{PV-TP}\end{bmatrix}\text{-}\begin{bmatrix}\text{BV}_j\\ \text{BC}_j\end{bmatrix}.$$

Theorem 3.3 *For each rational α, $0 \leq \alpha \leq 1$, and each choice from the independent choice brackets below, the specified parameterized (as j varies over $\mathbb{N}$) problem is in* FPT*:*

$$\begin{bmatrix}\text{succinct}\\ \emptyset\end{bmatrix}\text{-}\begin{bmatrix}\text{Copeland}^{\alpha}\\ \text{Copeland}^{\alpha}_{\text{Irrational}}\end{bmatrix}\text{-}\begin{bmatrix}\text{C}\\ \text{D}\end{bmatrix}\text{C}\begin{bmatrix}\text{AC}_{\text{u}}\\ \text{AC}\\ \text{DC}\\ \text{PC-TE}\\ \text{PC-TP}\\ \text{RPC-TE}\\ \text{RPC-TP}\end{bmatrix}\text{-BC}_j.$$

The proofs of Theorems 3.2 and 3.3, which in particular employ Lenstra's [29] algorithm for bounded-variable-cardinality integer programming, are omitted here.

FPT and Extended Control. We now introduce and look at extended control. By that we do not mean changing the basic control notions of adding/deleting/partitioning candidates/voters. Rather, we mean generalizing past merely looking at the constructive (make a distinguished candidate a winner) and the destructive (prevent a distinguished candidate from being a winner) cases. In particular, we are interested in control where the goal can be far more flexibly specified, for example (though in the partition cases we will be even more flexible than this), we will allow as our goal region any (reasonable—there are some time-related conditions) subcollection of "Copeland output tables" (specifications of who won/lost/tied each head-to-head contest).

Since from a Copeland output table, in concert with the current α, one can read off the Copeland$^{\alpha}_{\text{Irrational}}$ scores of the candidates, this allows us a tremendous range of descriptive flexibility in specifying our control goals, e.g., we can specify a linear order desired for the

candidates with respect to their Copeland$^{\alpha}_{\text{Irrational}}$ scores, we can specify a linear-order-with-ties desired for the candidates with respect to their Copeland$^{\alpha}_{\text{Irrational}}$ scores, we can specify the exact desired Copeland$^{\alpha}_{\text{Irrational}}$ scores for one or more candidates, we can specify that we want to ensure that no candidate from a certain subgroup has a Copeland$^{\alpha}_{\text{Irrational}}$ score that ties or beats the Copeland$^{\alpha}_{\text{Irrational}}$ score of any candidate from a certain other subgroup, etc.

All the FPT algorithms given in Theorems 3.2 and 3.3 regard, on their surface, the standard control problem, which tests whether a given candidate can be made a winner (constructive case) or can be precluded from being a winner (destructive case). We note that the general approaches used to prove those results in fact yield FPT schemes even for the far more flexible notions of control that we just mentioned.

Resistance Results. In contrast with the FPT results in Theorems 3.2 and 3.3, we show that for each rational α, $0 \leq \alpha \leq 1$, for Copeland$^{\alpha}_{\text{Irrational}}$ all the candidate-control cases that we showed earlier in this paper (i.e., without bounds on the number of voters) to be resistant remain resistant even for the case of bounded voters. This resistance holds even when the input is not in succinct format, and so it certainly also holds when the input is in succinct format.

It remains open whether Table 1's resistant, rational-voter, candidate-control cases remain resistant for the bounded-voter case.

4 Conclusions

We have shown that from the computational point of view the election systems of Llull and Copeland (i.e., Copeland$^{0.5}$) are broadly resistant to bribery and procedural control, regardless of whether the voters are required to have rational preferences. It is rather charming that Llull's 700-year-old system shows perfect resistance to bribery and more resistances to (constructive) control than any other natural system (even far more modern ones) with an easy winner-determination procedure—other than Copeland$^{\alpha}$, $0 < \alpha < 1$—is known to possess, and this is even more remarkable when one considers that Llull's system was defined long before control of elections was even explicitly studied. Copeland voting matches Llull's perfect resistance to bribery and in addition has perfect resistance to (constructive) control.

A natural open direction would be to study the complexity of control for additional election systems. Particularly interesting would be to seek existing, natural voting systems that have polynomial-time winner determination procedures but that are resistant to all standard types of both constructive *and destructive* control. Also extremely interesting would be to find single results that classify, for broad families of election systems, precisely what it is that makes control easy or hard, i.e., to obtain dichotomy meta-results for control (see Hemaspaandra and Hemaspaandra [25] for some discussion regarding work of that flavor for manipulation).

Acknowledgments: We thank Felix Brandt for pointing us to the work of Zermelo, and we thank the anonymous AAAI-07, AAIM-08, and COMSOC-08 referees for their helpful comments on earlier versions of this paper.

References

[1] A. Altman and M. Tennenholtz. An axiomatic approach to personalized ranking systems. In *Proc. IJCAI'07*, pages 1187–1192. AAAI Press, 2007.

[2] K. Arrow. *Social Choice and Individual Values*. John Wiley and Sons, 1951 (revised editon, 1963).

[3] J. Bartholdi, III and J. Orlin. Single transferable vote resists strategic voting. *Social Choice and Welfare*, 8(4):341–354, 1991.

[4] J. Bartholdi, III, C. Tovey, and M. Trick. The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6(3):227–241, 1989.

[5] J. Bartholdi, III, C. Tovey, and M. Trick. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare*, 6(2):157–165, 1989.

[6] J. Bartholdi, III, C. Tovey, and M. Trick. How hard is it to control an election? *Mathematical and Computer Modeling*, 16(8/9):27–40, 1992.

[7] J.-A.-N. de Caritat, Marquis de Condorcet. *Essai sur l'Application de L'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix*. 1785. Facsimile reprint of original published in Paris, 1972, by the Imprimerie Royale.

[8] V. Conitzer and T. Sandholm. Universal voting protocol tweaks to make manipulation hard. In *Proc. IJCAI'03*, pages 781–788. Morgan Kaufmann, August 2003.

[9] V. Conitzer and T. Sandholm. Nonexistence of voting rules that are usually hard to manipulate. In *Proc. AAAI'06*, pages 627–634. AAAI Press, July 2006.

[10] V. Conitzer, T. Sandholm, and J. Lang. When are elections with few candidates hard to manipulate? *Journal of the ACM*, 54(3):Article 14, 2007.

[11] A. Copeland. A "reasonable" social welfare function. Mimeographed notes from a Seminar on Applications of Mathematics to the Social Sciences, University of Michigan, 1951.

[12] R. Downey and M. Fellows. *Parameterized Complexity*. Springer-Verlag, 1999.

[13] J. Duggan and T. Schwartz. Strategic manipulability without resoluteness or shared beliefs: Gibbard–Satterthwaite generalized. *Social Choice and Welfare*, 17(1):85–93, 2000.

[14] E. Elkind and H. Lipmaa. Small coalitions cannot manipulate voting. In *Proc. FC'05*, pages 285–297. Springer-Verlag LNCS #3570, February/March 2005.

[15] G. Erdélyi, L. Hemaspaandra, J. Rothe, and H. Spakowski. On approximating optimal weighted lobbying, and frequency of correctness versus average-case polynomial time. In *Proc. FCT'07*, pages 300–311. Springer-Verlag LNCS #4639, August 2007.

[16] G. Erdélyi, M. Nowak, and J. Rothe. Sincere-strategy preference-based approval voting broadly resists control. In *Proc. MFCS'08*, pages 311–322. Springer-Verlag LNCS #5162, August 2008.

[17] G. Erdélyi, M. Nowak, and J. Rothe. Sincere-strategy preference-based approval voting fully resists constructive control and broadly resists destructive control. Technical Report cs.GT/0806.0535, ACM Computing Research Repository (CoRR), June 2008. A precursor appears as [16].

[18] P. Faliszewski. Nonuniform bribery (short paper). In *Proc. AAMAS'08*, pages 1569–1572, May 2008.

[19] P. Faliszewski, E. Hemaspaandra, and L. Hemaspaandra. The complexity of bribery in elections. In *Proc. AAAI'06*, pages 641–646. AAAI Press, 2006.

[20] P. Faliszewski, E. Hemaspaandra, L. Hemaspaandra, and J. Rothe. Llull and Copeland voting broadly resist bribery and control. In *Proc. AAAI'07*, pages 724–730. AAAI Press, 2007.

[21] P. Faliszewski, E. Hemaspaandra, L. Hemaspaandra, and J. Rothe. Copeland voting fully resists constructive control. In *Proc. AAIM'08*, pages 165–176. Springer-Verlag, June 2008. To appear.

[22] P. Faliszewski, E. Hemaspaandra, and H. Schnoor. Copeland voting: Ties matter. In *Proc. AAMAS'08*, pages 983–990, May 2008.

[23] A. Gibbard. Manipulation of voting schemes. *Econometrica*, 41(4):587–601, 1973.

[24] G. Hägele and F. Pukelsheim. The electoral writings of Ramon Llull. *Studia Lulliana*, 41(97):3–38, 2001.

[25] E. Hemaspaandra and L. Hemaspaandra. Dichotomy for voting systems. *Journal of Computer and System Sciences*, 73(1):73–83, 2007.

[26] E. Hemaspaandra, L. Hemaspaandra, and J. Rothe. Anyone but him: The complexity of precluding an alternative. *Artificial Intelligence*, 171(5-6):255–285, April 2007.

[27] E. Hemaspaandra, L. Hemaspaandra, and J. Rothe. Hybrid elections broaden complexity-theoretic resistance to control. In *Proc. IJCAI'07*, pages 1308–1314. AAAI Press, 2007.

[28] C. Homan and L. Hemaspaandra. Guarantees for the success frequency of an algorithm for finding Dodgson-election winners. *Journal of Heuristics*. To appear. Full version available as Department of Computer Science, University of Rochester Technical Report TR-881, September 2005, revised June 2007.

[29] H. Lenstra, Jr. Integer programming with a fixed number of variables. *Mathematics of Operations Research*, 8(4):538–548, 1983.

[30] J. McCabe-Dansted, G. Pritchard, and A. Slinko. Approximability of Dodgson's rule. *Social Choice and Welfare*, 2008. DOI 10.1007/s00355-007-0282-8.

[31] I. McLean and A. Urken. *Classics of Social Choice*. University of Michigan Press, 1995.

[32] V. Merlin and D. Saari. Copeland method II: Manipulation, monotonicity, and paradoxes. *Journal of Economic Theory*, 72(1):148–172, 1997.

[33] A. Procaccia and J. Rosenschein. Junta distributions and the average-case complexity of manipulating elections. *Journal of Artificial Intelligence Research*, 28:157–181, February 2007.

[34] A. Procaccia, J. Rosenschein, and G. Kaminka. On the robustness of preference aggregation in noisy environments. In *Proc. AAMAS'07*, pages 416–422. ACM Press, 2007.

[35] A. Procaccia, J. Rosenschein, and A. Zohar. Multi-winner elections: Complexity of manipulation, control, and winner-determination. In *Proc. IJCAI'07*, pages 1476–1481. AAAI Press, 2007.

[36] D. Saari and V. Merlin. The Copeland method I: Relationships and the dictionary. *Economic Theory*, 8(1):51–76, 1996.

[37] M. Satterthwaite. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10(2):187–217, 1975.

[38] E. Zermelo. Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29(1):436–460, 1929.

Piotr Faliszewski and Lane A. Hemaspaandra
Department of Computer Science
University of Rochester
Rochester, NY 14627, USA

Edith Hemaspaandra
Department of Computer Science
Rochester Institute of Technology
Rochester, NY 14623, USA

Jörg Rothe
Institut für Informatik
Heinrich-Heine-Universität Düsseldorf
40225 Düsseldorf, Germany

On Voting Caterpillars: Approximating Maximum Degree in a Tournament by Binary Trees

Felix Fischer, Ariel D. Procaccia and Alex Samorodnitsky

Abstract

Voting trees describe an iterative procedure for selecting a single vertex from a tournament. It has long been known that there is no voting tree that always singles out a vertex with maximum degree. In this paper, we study the power of voting trees in *approximating* the maximum degree. We give upper and lower bounds on the worst-case ratio between the degree of the vertex chosen by a tree and the maximum degree, both for the deterministic model concerned with a single fixed tree, and for randomizations over arbitrary sets of trees. Our main positive result is a randomization over surjective trees of polynomial size that provides an approximation ratio of at least 1/2. The proof is based on a connection between a randomization over caterpillar trees and a rapidly mixing Markov chain.

1 Introduction

A problem that pervades the theory of social choice is the selection of "best" alternatives from a *tournament*, *i.e.*, a complete and asymmetric (dominance) relation over a set of alternatives (see, *e.g.*, Laslier, 1997). Such a relation for example arises from pairwise majority voting with an odd number of voters and linear preferences. In graph theoretic terms, a tournament is an orientation of a complete undirected graph, with a directed edge from a dominating alternative to a dominated one. In the presence of cycles the concept of maximality is not well-defined, and so-called tournament solutions have been devised to take over the role of singling out good alternatives. A prominent such solution, known as the Copeland solution, selects the alternatives with *maximum (out-)degree*, *i.e.*, those that beat the largest number of other alternatives in a direct comparison.

An interesting question concerns the implementation of a solution concept using a specific procedure. We shall specifically be interested in the well-known class of procedures given by *voting trees*. A voting tree over a set A of alternatives is a binary tree with leaves labeled by elements of A. Given a tournament T, a labeling for the internal nodes is defined recursively by labeling a node by the label of its child that beats the other child according to T (or by the unique label of its children if both have the same label). The label at the root is then deemed the winner of the voting tree given tournament T. This definition expressly allows an alternative to appear multiple times at the leaves of a tree.

A voting tree over A is said to *implement* a particular solution concept if for every tournament on A it selects an optimal alternative according to said solution concept. It has long been known that there exists no voting tree implementing the Copeland solution, *i.e.*, one that always selects a vertex with maximum degree (Moulin, 1986). In this paper, we ask a natural question from a computer science point of view: Is there a voting tree that *approximates* the maximum degree? More precisely, we would like to determine the largest value of α, such that for any set A of alternatives, there exists a tree Γ, which for every tournament on A selects an alternative with at least α times the maximum degree in the tournament. We will address this question both in the *deterministic* model, where Γ is a fixed voting tree, and in the *randomized* model, where voting trees are chosen randomly

according to some distribution.

Results Our main negative results are upper bounds of 3/4 and 5/6, respectively, on the approximation ratio achievable by deterministic trees and randomizations over trees. We find it quite surprising that randomizations over trees cannot achieve a ratio arbitrarily close to 1.

For most of the paper we concentrate on the randomized model. We study a class of trees we call voting caterpillars, which are characterized by the fact that they have exactly two nodes on each level below the root. We devise a randomization over "small" trees of this type, which further satisfies an important property we call *admissibility*: its support only contains trees where every alternative appears in some leaf. Our main positive result is the following.

Theorem 4.1. *Let A be a set of alternatives. Then there exists an admissible randomization over voting trees on A of size polynomial in $|A|$ with an approximation ratio of $1/2 - \mathcal{O}(1/|A|)$.*

We prove this theorem by establishing a connection to a nonreversible, rapidly mixing random walk on the tournament, and analyzing its stationary distribution. The proof of rapid mixing involves reversibilizing the transition matrix, and then bounding its spectral gap via its conductance. We further show that our analysis is tight, and that voting caterpillars also provide a lower bound of 1/2 for the second order degree of an alternative, defined as the sum of degrees of those alternatives it dominates.

The paper concludes with negative results about more complex tree structures, which turn out to be rather surprising. In particular, we show that the approximation ratio provided by randomized balanced trees can become arbitrarily bad with growing height. We further show that "higher-order" caterpillars, with labels chosen by lower-order caterpillars instead of uniformly at random, can also cause the approximation ratio to deteriorate.

Related Work In economics, the problem of implementation by voting trees was introduced by Farquharson (1969), and further explored, for example, by McKelvey and Niemi (1978), Miller (1980), Moulin (1986), Herrero and Srivastava (1992), Dutta and Sen (1993), Srivastava and Trick (1996), and Coughlan and Le Breton (1999). In particular, Moulin (1986) shows that the Copeland solution is not implementable by voting trees if there are at least 8 alternatives, while Srivastava and Trick (1996) demonstrate that it can be implemented for tournaments with up to 7 alternatives.

Laffond et al. (1994) compute the *Copeland measure* of several prominent *choice correspondences*—functions mapping each tournament to a *set* of desirable alternatives. In contrast to the (Copeland) approximation ratio considered in this paper, the Copeland measure is computed with respect to the best alternative selected by the correspondence, so strictly speaking it is not a worst-case measure. More importantly, however, Laffond et al. study properties of given correspondences, whereas we investigate the possibility of *constructing* voting trees with certain desirable properties. In this sense, our work is algorithmic in nature, while theirs is descriptive.

In theoretical computer science, the problem studied in this paper is somewhat reminiscent of the problem of determining query complexity of graph properties (see, *e.g.*, Rosenberg, 1973; Rivest and Vuillemin, 1976; Kahn et al., 1984; King, 1988). In the general model, one is given an unknown graph over a known set of vertices, and must determine whether the graph satisfies a certain property by querying the edges. The complexity of a property is then defined as the height of the smallest decision tree that checks the property. Voting trees can be interpreted as querying the edges of the tournament in parallel, and in

a way that severely limits the ways in which, and the extent up to which, information can be transferred between different queries.

In the area of computational social choice, which lies at the boundary of computer science and economics, several authors have looked at the computational properties of voting trees and of various solution concepts. For example, Lang et al. (2007) characterize the computational complexity of determining different types of winners in voting trees. Procaccia et al. (2007) investigate the learnability of voting trees, as functions from tournaments to alternatives. In a slightly different context, Brandt et al. (2007) study the computational complexity of different solution concepts, including the Copeland solution.

Organization We begin by introducing the necessary concepts and notation. In Section 3 we present upper bounds for the deterministic and the randomized setting. In Section 4, we establish our main positive result using a randomization over voting caterpillars. Section 5 is devoted to balanced trees, and Section 6 concludes with some open questions. Proofs of all results, as well as an analysis of "higher order" caterpillars, can be found in the full version of this paper (Fischer et al., 2008).

2 Preliminaries

Let $A = \{1, \dots, m\}$ be a set of *alternatives*. A *tournament* T on A is an orientation of the complete graph with vertex set A. We denote by $\mathcal{T}(A)$ the set of all tournaments on A. For a tournament $T \in \mathcal{T}(A)$, we write iTj if the edge between a pair $i, j \in A$ of alternatives is directed from i to j, or i *dominates* j. For an alternative $i \in A$ we denote by $s_i = s_i(T) = |\{j \in A : iTj\}|$ the *degree* or (Copeland) *score* of i, *i.e.*, the number of outgoing edges from this alternative, omitting T when it is clear from the context.

We then consider computations performed by a specific type of tree on a tournament. In the context of this paper, a (deterministic) *voting tree* on A is a structure $\Gamma = (V, E, \ell)$ where (V, E) is a binary tree with root $r \in V$, and $\ell : V \to A$ is a mapping that assigns an element of A to each leaf of (V, E). Given a tournament T, a unique function $\ell_T : V \to A$ exists such that

$$\ell_T(v) = \begin{cases} \ell(v) & \text{if } v \text{ is a leaf} \\ \ell(u_1) & \text{if } v \text{ has children } u_1 \text{ and } u_2, \text{ and } \ell(u_1)T\ell(u_2) \text{ or } \ell(u_1) = \ell(u_2) \end{cases}$$

We will be interested in the label of the root r under this labeling, which we call the winner of the tree and denote by $\Gamma(T) = \ell_T(r)$. We call a tree Γ *surjective* if ℓ is surjective. Obviously, surjectivity corresponds to a very basic fairness requirement on the solution implemented by a tree. Other authors therefore view surjectivity as an inherent property of voting trees and define them accordingly (see, *e.g.*, Moulin, 1986). The sole reason we do not require surjectivity by definition is that our analysis will use trees that are not necessarily surjective.

Finally, a voting tree Γ on A will be said to provide an approximation ratio of α (w.r.t. the maximum degree) if

$$\min_{T \in \mathcal{T}(A)} \frac{s_{\Gamma(T)}}{\max_{i \in A} s_i(T)} \geq \alpha.$$

The above model can be generalized by looking at *randomizations* over voting trees according to some probability distribution. We will call a randomization *admissible* if its support contains only surjective trees. A distribution Δ over voting trees will then be said to provide a (randomized) approximation ratio of α if

$$\min_{T \in \mathcal{T}(A)} \frac{\mathbb{E}_{\Gamma \sim \Delta}[s_{\Gamma(T)}]}{\max_{i \in A} s_i(T)} \geq \alpha.$$

While we are of course interested in the approximation ratio achievable by admissible randomizations, it will prove useful to consider a specific class of randomizations that are not admissible, namely those that choose uniformly from the set of all voting trees with a given structure. Equivalently, such a randomization is obtained by fixing a binary tree and assigning alternatives to the leaves independently and uniformly at random, and will thus be called a *randomized voting tree*.

3 Upper Bounds

In this section we derive upper bounds on the approximation ratio achievable by voting trees, both in the deterministic model and in the randomized model. We build on concepts and techniques introduced by Moulin (1986), and begin by quickly familiarizing the reader with these.

Given a tournament T on a set A of alternatives, we say that $C \subseteq A$ is a *component*[1] of T if for all $i_1, i_2 \in C$ and $j \in A \setminus C$, $i_1 T j$ if and only if $i_2 T j$. For a component C, denote by $\mathcal{T}_C$ the subset of tournaments that have C as a component. If $T \in \mathcal{T}_C$, we can unambiguously define a tournament T_C on $(A \setminus C) \cup \{C\}$ by replacing the component C by a single alternative. The following lemma states that for two tournaments that differ only inside a particular component, any tree chooses an alternative from that component for one of the tournaments if and only if it does for the other. Furthermore, if an alternative outside the component is chosen for one tournament, then the same alternative has to be chosen for the other. Laslier (1997) calls a solution concept satisfying these properties *weakly composition-consistent*.

Lemma 3.1 (Moulin 1986)**.** *Let A be a set of alternatives, Γ a voting tree on A. Then, for all proper subsets $C \subsetneq A$, and for all $T, T' \in \mathcal{T}_C$,*

1. *$[T_C = T'_C]$ implies $[\Gamma(T) \in C$ if and only if $\Gamma(T') \in C]$, and*
2. *$[T_C = T'_C$ and $\Gamma(T) \in A \setminus C]$ implies $[\Gamma(T) = \Gamma(T')]$.*

We are now ready to strengthen the negative result concerning implementability of the Copeland solution (Moulin, 1986) by showing that no deterministic tree can always choose an alternative that has a degree significantly larger than 3/4 of the maximum degree.

Theorem 3.2. *Let A be a set of alternatives, $|A| = m$, and let Γ be a deterministic voting tree on A with approximation ratio α. Then, $\alpha \leq 3/4 + \mathcal{O}(1/m)$.*

Proof. For ease of exposition, we assume $|A| = m = 3k + 1$ for some odd k, but the same result (up to lower order terms) holds for all values of m. Define a tournament T comprised of three components C_1, C_2, and C_3, such that for $r = 1, 2, 3$, (i) $|C_r| = k$ and the restriction of T to C_r is regular, *i.e.*, each $i \in C_r$ dominates exactly $(k-1)/2$ of the alternatives in C_r, and (ii) for all $i \in C_r$ and $j \in C_{(r \bmod 3)+1}$, iTj. An illustration for $k = 3$ is given on the left of Figure 1.

Now consider any deterministic voting tree Γ on A, and assume w.l.o.g. that $\Gamma(T) \in C_1$. Define T' to be a tournament on A such that the restrictions of T and T' to $B \subseteq A$ are identical if $|B \cap C_2| \leq 1$, and the restriction of T' to C_2 is transitive; in particular, there is $i \in C_2$ such that for any $i \neq j \in C_2$, $iT'j$. An illustration for $k = 3$ is given on the right of Figure 1. By Lemma 3.1, $\Gamma(T') = \Gamma(T)$. Furthermore, T' satisfies

$$s_{\Gamma(T')} = k + \frac{(k-1)}{2} = \frac{3k}{2} - \frac{1}{2} \qquad \text{and} \qquad \max_{i \in A} s_i = 2k - 1,$$

[1] Moulin (1986) uses the term "adjacent set".

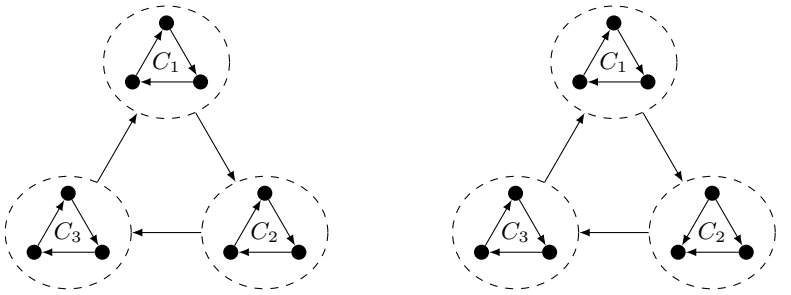


Figure 1: Tournaments used in the proof of Theorem 3.2, illustrated for $k = 3$. A voting tree is assumed to select an alternative from C_1.

and thus

$$\frac{s_{\Gamma(T')}}{\max_{i \in A} s_i(T')} = \frac{3k-1}{4k-2} \leq \frac{3(k-1)+2}{4(k-1)} = \frac{3}{4} + \frac{1}{2(k-1)}.$$

□

We now turn to the randomized model. It turns out that one cannot obtain an approximation ratio arbitrarily close to 1 by randomizing over large trees. We derive an upper bound for the approximation ratio by using similar arguments as in the deterministic case above, and combining them with the minimax principle of Yao (1977).

Theorem 3.3. *Let A be a set of alternatives, $|A| = m$, and let Δ be a probability distribution over voting trees on A with an approximation ratio of α. Then, $\alpha \leq 5/6 + \mathcal{O}(1/m)$.*

The proof of this theorem is given in the full version of the paper. We point out that the theorem holds in particular for inadmissible randomizations.

4 A Randomized Lower Bound

A weak deterministic lower bound of $\Theta((\log m)/m)$ can be obtained straightforwardly from a balanced tree where every label appears exactly once. While balanced trees will be discussed in more detail in Section 5, they become increasingly unwieldy with growing height, and an improvement of this lower bound or of the deterministic upper bound given in the previous section currently seems to be out of our reach. In the remainder of the paper, we therefore concentrate on the randomized model.

In this section we put forward our main result, a lower bound of $1/2$, up to lower order terms, for admissible randomizations over voting trees. Let us state the result formally.

Theorem 4.1. *Let A be a set of alternatives. Then there exists an admissible randomization over voting trees on A of size polynomial in $|A|$ with an approximation ratio of $1/2 - \mathcal{O}(1/m)$.*

In addition to satisfying the basic admissibility requirement, the randomization also has the desirable property of relying only on trees of polynomial size. This clearly facilitates its use as a computational procedure. To prove Theorem 4.1, we make use of a specific binary tree structure known as caterpillar trees.

4.1 Randomized Voting Caterpillars

We begin by inductively defining a family of binary trees that we refer to as *k-caterpillars*. The 1-caterpillar consists of a single leaf. A k-caterpillar is a binary tree, where one subtree of the root is a $(k-1)$-caterpillar, and the other subtree is a leaf. Then, a *voting k-caterpillar* on A is a k-caterpillar whose leaves are labeled by elements of A.

It is straightforward to see that an upper and lower bound of $1/2$ holds for the randomized 1-caterpillar, *i.e.*, the uniform distribution over the m possible voting 1-caterpillars. Indeed, such a tree is equivalent to selecting an alternative uniformly at random. Since we have $\sum_{i \in A} s_i = \binom{m}{2}$, the expected score of a random alternative is $(m-1)/2$, whereas the maximum possible score is $m-1$. This randomization, however, like other randomizations over small trees that conceivably provide a good approximation ratio, is not admissible and actually puts probability one on trees that are not surjective.

To prove Theorem 4.1, we instead use the uniform randomization over surjective k-caterpillars, henceforth denoted k-RSC, which is clearly admissible. Theorem 4.1 can then be restated as a more explicit—and slightly stronger—result about the k-RSC.

Lemma 4.2. *Let A be a set of alternatives, $T \in \mathcal{T}(A)$. For $k \in \mathbb{N}$, denote by $p_i^{(k)}$ the probability that alternative $i \in A$ is selected from T by the k-RSC. Then, for every $\epsilon > 0$ there exists $k = k(m, \epsilon)$ polynomial in m and $1/\epsilon$ such that*

$$\sum_{i \in A} p_i^{(k)} s_i \geq \frac{m-1}{2} - \epsilon.$$

To see that this directly implies Theorem 4.1, recall that the maximum score is $m-1$ and choose, for example, $\epsilon = 1$. The remainder of this section is devoted to the proof of Lemma 4.2. For the sake of analysis, we will use the randomized k-caterpillar, or k-RC, as a proxy to the k-RSC. We recall that the k-RC is equivalent to a k-caterpillar with labels for the leaves chosen independently and uniformly at random. In other words, it corresponds to the uniform distribution over all possible voting k-caterpillars, rather than just the surjective ones.

Clearly the k-RC corresponds to a randomization that is not admissible. In contrast to very small trees, however, like the one consisting only of a single leaf, it is straightforward to show that the distribution over alternatives selected by the RC is very close to that of the RSC.

Lemma 4.3. *Let $k \geq m$, and denote by $\bar{p}_i^{(k)}$ and $p_i^{(k)}$, respectively, the probability that alternative $i \in A$ is selected by the k-RC and by the k-RSC for some tournament $T \in \mathcal{T}(A)$. Then, for all $i \in A$,*

$$|\bar{p}_i^{(k)} - p_i^{(k)}| \leq \frac{m}{e^{k/m}}.$$

Proof. For all $i \in A$, $|\bar{p}_i^{(k)} - p_i^{(k)}|$ is at most the probability that the k-RC does not choose a surjective tree. By the union bound, we can bound this probability by

$$\sum_{i \in A} \Pr[i \text{ does not appear in the } k\text{-RC}] \leq m \cdot \left(1 - \frac{1}{m}\right)^k \leq \frac{m}{e^{k/m}}.$$

□

With Lemma 4.3 at hand, we can temporarily restrict our attention to the k-RC. A direct analysis of the k-RC, and in particular of the competition between the winner of the $(k-1)$-RC and a random alternative, shows that for every k, the k-RC provides an approximation

ratio of at least 1/3. It seems, however, that this analysis cannot be extended to obtain an approximation ratio of 1/2. In order to reach a ratio of 1/2, we shall therefore proceed by employing a second abstraction. Given a tournament T, we define a Markov chain $\mathfrak{M} = \mathfrak{M}(T)$ as follows:[2] The state space Ω of $\mathfrak{M}$ is A, and its initial distribution $\pi^{(0)}$ is the uniform distribution over Ω. The transition matrix $P = P(T)$ is given by

$$P(i,j) = \begin{cases} \frac{s_i+1}{m} & \text{if } i = j \\ \frac{1}{m} & \text{if } jTi \\ 0 & \text{if } iTj. \end{cases}$$

We claim that the distribution $\pi^{(k)}$ of $\mathfrak{M}$ after k steps is exactly the probability distribution $\bar{p}^{(k+1)}$ over alternatives selected by the $(k+1)$-RC. In order to see this, note that the 1-RC chooses an alternative uniformly at random. Then, the winner of the k-RC is the winner of the $(k-1)$-RC if the latter dominates, or is identical to, the alternative assigned to the other child of the root. This happens with probability $(s_i+1)/m$ when i is the winner of the k-RC. Otherwise the winner is some other alternative that dominates the winner of the k-RC, and each such alternative is assigned to the other child of the root with probability $1/m$.

We shall be interested in the performance guarantees given by the stationary distribution π of $\mathfrak{M}$. We first show that $\mathfrak{M}$ is guaranteed to converge to a unique such distribution, despite the fact that it is not necessarily irreducible.

Lemma 4.4. *Let T be a tournament. Then $\mathfrak{M}(T)$ converges to a unique stationary distribution.*

The proof of the lemma appears in the full version of the paper. We are now ready to show that an alternative drawn from the stationary distribution will have an expected degree of at least half the maximum possible degree.

Lemma 4.5. *Let $T \in \mathcal{T}(A)$ be a tournament, π the stationary distribution of $\mathfrak{M}(T)$. Then*

$$\sum_{i \in A} \pi_i s_i \geq \frac{m-1}{2}.$$

The proof is based on some algebraic manipulations and the Cauchy-Schwarz inequality, and is delegated to the full version of the paper.

The last ingredient in the proof of Lemma 4.2 and Theorem 4.1 is to show that for some k polynomial in m, the distribution over alternatives selected by the k-RC, which we recall to be equal to the distribution of $\mathfrak{M}$ after $k-1$ steps, is close to the stationary distribution of $\mathfrak{M}$. In other words, we want to show that for every tournament T, $\mathfrak{M}(T)$ is rapidly mixing.[3]

Lemma 4.6. *Let T be a tournament. Then, for every $\epsilon > 0$ there exists $k = k(m, \epsilon)$ polynomial in m and $1/\epsilon$, such that for all $k' > k$ and all $i \in A$, $|\pi_i^{(k')} - \pi_i| \leq \epsilon$, where $\pi^{(k)}$ is the distribution of $\mathfrak{M}(T)$ after k steps and π is the stationary distribution of $\mathfrak{M}(T)$.*

The proof of Lemma 4.6, which is given in the full version of the paper, works by reversibilizing the transition matrix of $\mathfrak{M}$ and then bounding the spectral gap of the reversibilized matrix via its conductance.

We now have all the necessary ingredients in place.

[2] Curiously, this chain bears resemblance to one previously used to define a solution concept called the Markov set (see, *e.g.*, Laslier, 1997). However, only limited attention has been given to a formal analysis of this chain, concerning properties which are different from the ones we are interested in.

[3] We might be slightly abusing terminology here, since the theory of rapidly mixing Markov chains usually considers chains with an exponential state space, which converge in time poly-logarithmic in the size of the state space. In our case the size of the state space is only m, and the mixing rate is polynomial in m.

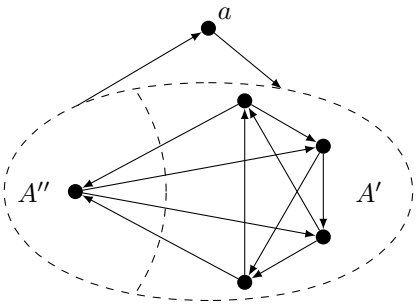


Figure 2: Tournament structure providing an upper bound for the randomized k-caterpillar, example for $m = 6$ and $\epsilon = 1/5$. A' and A'' contain $(1-\epsilon)(m-1)$ and $\epsilon(m-1)$ alternatives, respectively.

Proof of Lemma 4.2 and Theorem 4.1. Let $\epsilon > 0$. By Lemma 4.3 and Lemma 4.6, there exists k polynomial in m and $1/\epsilon$ such that for all $i \in A$, $|p_i^{(k)} - \bar{p}_i^{(k)}| \leq \epsilon/(2\binom{m}{2})$ and $|\bar{p}_i^{(k)} - \pi_i| \leq \epsilon/(2\binom{m}{2})$. By the triangle inequality, $|p_i^{(k)} - \pi_i| \leq \epsilon/\binom{m}{2}$. Now,

$$\sum_i \pi_i s_i - \sum_i p_i^{(k)} s_i \leq \sum_i |\pi_i - p_i^{(k)}| s_i \leq \frac{\epsilon}{\binom{m}{2}} \sum_i s_i = \epsilon.$$

Lemma 4.2 and thus Theorem 4.1 follow directly by Lemma 4.5. □

4.2 Tightness and Stability of the Caterpillar

It turns out that the analysis in the proof of Theorem 4.1 is tight. Indeed, since we have seen that the stationary distribution π of $\mathfrak{M}$ is very close to the distribution of alternatives chosen by the k-RSC, it is sufficient to see that π cannot guarantee an approximation ratio better than $1/2$ in expectation. Consider a set A of alternatives, and a partition of A into three sets A', A'', and $\{a\}$ such that $|A'| = (1-\epsilon)(m-1)$ and $|A''| = \epsilon(m-1)$ for some $\epsilon > 0$. Further consider a tournament $T \in \mathcal{T}(A)$ in which a dominates every alternative in A' and is itself dominated by every alternative in A'', and for which the restriction of T to $A' \cup A''$ is regular. The structure of T is illustrated in Figure 2.

It is easily verified that the stationary distribution π of $\mathfrak{M}(T)$ satisfies

$$\pi_a = \frac{\sum_{j:aTj} \pi_j}{m - s_a - 1} \leq \frac{1}{m - s_a - 1} \leq \frac{1}{\epsilon(m-1)},$$

and therefore,

$$\sum_i \pi_i s_i \leq \frac{1}{\epsilon(m-1)}(m-1) + \frac{\epsilon(m-1)-1}{\epsilon(m-1)} \cdot \left(\frac{m-1}{2} + 1\right) \leq \frac{m-1}{2} + \frac{1}{\epsilon} + 1.$$

Furthermore, a has degree $(1-\epsilon)(m-1)$. If we choose, say, $\epsilon = 1/\sqrt{m}$, then the approximation ratio tends to $1/2$ as m tends to infinity.

We proceed to demonstrate that the above tournament is a generic bad example. Indeed, Lemma 4.5 will be shown to possess the following stability property: in every tournament where π achieves an approximation ratio only slightly better than $1/2$, almost all alternatives have degree close to $m/2$, as it is the case for the example above. In particular, this implies that $\mathfrak{M}$ either provides an expected approximation ratio better than $1/2$, or selects an alternative with score around $m/2$ with very high probability.

Theorem 4.7. *Let $\epsilon > 0$, $m \geq 1/(2\sqrt{\epsilon})$. Let T be a tournament over a set of m alternatives, π the stationary distribution of $\mathfrak{M}(T)$. If $\sum_i \pi_i s_i = (m-1)/2 + \epsilon m$, then*

$$\left| \left\{ i \in A : \left| s_i - \frac{m}{2} \right| > \frac{3\sqrt[4]{4\epsilon}}{2} m \right\} \right| \leq \sqrt[4]{4\epsilon} \cdot m.$$

The details of the proof appear in the full version of the paper.

4.3 Second Order Degrees

So far we have been concerned with the Copeland solution, which selects an alternative with maximum degree. Recently, a related solution concept, sometimes referred to as *second order Copeland*, has received attention in the social choice literature (see, *e.g.*, Bartholdi et al., 1989). Given a tournament T, this solution breaks ties with respect to the maximum degree toward alternatives i with maximum *second order degree* $\sum_{j:iTj} s_j$. Second order Copeland is the first rule, and one of only two natural voting rules, known to be computationally easy to compute but difficult to manipulate (Bartholdi et al., 1989).

Interestingly, the same randomization studied in Section 4.1 also achieves a 1/2-approximation for the second order degree.

Theorem 4.8. *Let A be a set of alternatives, $T \in \mathcal{T}(A)$. For $k \in \mathbb{N}$, let $p_i^{(k)}$ denote the probability that alternative $i \in A$ is selected by the k-RSC for T. Then, there exists $k = k(m)$ polynomial in m such that*

$$\frac{\sum p_i^{(k)} \sum_{j:iTj} s_j}{\max_{i \in A} \sum_{j:iTj} s_j} \geq \frac{1}{2} + \Omega(1/m).$$

Clearly, the sum of degrees of alternatives dominated by an alternative i is at most $\binom{m-1}{2}$. The lower bound is then obtained from an explicit result about the second order degree of alternatives chosen by the k-RSC. Along similar lines as in the proof of Theorem 4.1, it suffices to prove that the stationary distribution of $\mathfrak{M}(T)$ provides an approximation. The following lemma is the second order analog of Lemma 4.5.

Lemma 4.9. *Let T be a tournament, π the stationary distribution of $\mathfrak{M}(T)$. Then,*

$$\sum_{i \in A} \left(\pi_i \sum_{j:iTj} s_j \right) \geq \frac{m^2}{4} - \frac{m}{2}.$$

It turns out that the technique used in the proof of Lemma 4.5, namely directly manipulating the stationary distribution equations and applying Cauchy-Schwarz, does not work for the second order degree. We instead formulate a suitable LP and bound the primal by a feasible solution to the dual. The proof of the lemma, which in turn implies Theorem 4.8, is delegated to the full version of the paper.

We further point out that the analysis is tight. Indeed, the second order degree of any alternative in a regular tournament, *i.e.*, one where each alternative dominates exactly $(m-1)/2$ other alternatives, is $(m-1)/2 \cdot (m-1)/2 = m^2/4 - m/2 + 1/4$. Theorem 4.8 itself is also tight, by the example given in Section 4.2.

5 Balanced Trees

In the previous section we presented our main positive results, all of which were obtained using randomizations over caterpillars. Since caterpillars are maximally unbalanced, one

would hope to do much better by looking at *balanced trees*, *i.e.*, trees where the depth of any two leaves differs by at most one. We briefly explore this intuition. Consider a balanced binary tree where each alternative in a set A appears exactly once at a leaf. We will call such a tree a *permutation tree* on A. As we have already mentioned in the previous section, permutations trees provide a very weak deterministic lower bound. Indeed, the winning alternative must dominate the $\Theta(\log m)$ alternatives it meets on the path to the root, all of which are distinct. Since there always exists an alternative with score at least $(m-1)/2$, we obtain an approximation ratio of $\Theta((\log m)/m)$. On the other hand, no voting tree in which every two leaves have distinct labels can guarantee to choose an alternative with degree larger than the height of the tree, so the above bound is tight. More interestingly, it can be shown that no composition of permutation trees, *i.e.*, no tree obtained by replacing every leaf of an arbitrary binary tree by a permutation tree, can provide a lower bound better than $1/2$. Unfortunately, larger balanced trees not built from permutation trees have so far remained elusive.

Can we obtain a better bound by randomizing? Intuitively, a randomization over large balanced trees should work well, because one would expect that the winning alternative dominate a large number of randomly chosen alternatives on the way to the root. Surprisingly, the complete opposite is the case. In the following, we call *randomized perfect voting tree* of height k, or k-RPT, a voting tree where every leaf is at depth k and labels are assigned uniformly at random. This tree obviously corresponds to a randomization that is not admissible, but a similar result for admissible randomizations can easily be obtained by using the same arguments as before.

Theorem 5.1. *Let A be a set of alternatives, $|A| \geq 5$. For every $K \in \mathbb{N}$ and $\epsilon > 0$, there exists $K' \geq K$ such that the K'-RPT provides an approximation ratio of at most $\mathcal{O}(1/m)$.*

The proof of this theorem, which is given in the full version of the paper, constructs a tournament consisting of a 3-cycle of components and shows that the distribution over alternatives chosen by the k-RPT *oscillates* between the different components as k grows.

We have analyzed higher order voting caterpillars obtained by replacing each leaf of a caterpillar of sufficiently large height by higher order caterpillars of smaller order (in particular, of order reduced by one). As in the case of the k-RPT, this construction does not provide better bounds but instead causes the approximation ratio to deteriorate.

6 Open Problems

Many interesting questions arise from our work. Perhaps the most enigmatic open problem in the context of this paper concerns tighter bounds for deterministic trees. Some results for restricted classes of trees have been discussed in Section 5, but in general there remains a large gap between the upper bound of $3/4$ derived in Section 3 and the straightforward lower bound of $\Theta((\log m)/m)$.

In the randomized model our situation is somewhat better. Nevertheless, an intriguing gap remains between our upper bound of $5/6$, which holds even for inadmissible randomizations over arbitrarily large trees, and the lower bound of $1/2$ obtained from an admissible randomization over trees of polynomial size. It might be the case that the height of a k-RPT could be chosen carefully to obtain some kind of approximation guarantee. For example, one could investigate the uniform distribution over permutation trees. The analysis of this type of randomization is closely related to the theory of dynamical systems, and we expect it to be rather involved.

Acknowledgements

We thank Julia Böttcher, Felix Brandt, Shahar Dobzinski, Dvir Falik, Jeff Rosenschein, and Michael Zuckerman for many helpful discussions.

Part of the work was done while Felix Fischer was visiting The Hebrew University of Jerusalem. This visit was supported by a Minerva Short-Term Research Grant. Ariel Procaccia is supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities. Alex Samorodnitsky is supported by ISF grant 039-7165.

References

J. Bartholdi, C. A. Tovey, and M. A. Trick. The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6:227–241, 1989.

F. Brandt, F. Fischer, and P. Harrenstein. The computational complexity of choice sets. In *Proceedings of the Eleventh Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pages 82–91, 2007.

P. J. Coughlan and M. Le Breton. A social choice function implementable via backward induction with values in the ultimate uncovered set. *Review of Economic Design*, 4: 153–160, 1999.

B. Dutta and A. Sen. Implementing generalized Condorcet social choice functions via backward induction. *Social Choice and Welfare*, 10:149–160, 1993.

R. Farquharson. *Theory of Voting*. Yale University Press, 1969.

F. Fischer, A. D. Procaccia, and A. Samorodnitsky. On voting caterpillars: Approximating maximum degree in a tournament by binary trees. Technical Report 2008-06, Leibniz Center for Research in Computer Science, Hebrew University, 2008. Available from http://www.cs.huji.ac.il/~arielpro/papers/rt.pdf.

M. Herrero and S. Srivastava. Decentralization by multistage voting procedures. *Journal of Economic Theory*, 56:182–201, 1992.

J. Kahn, M. Saks, and D. Sturtevant. A topological approach to evasiveness. *Combinatorica*, 4:297–306, 1984.

V. King. Lower bounds on the complexity of graph properties. In *Proceedings of the 20th ACM Symposium on Theory of Computing (STOC)*, pages 468–476, 1988.

G. Laffond, J. F. Laslier, and M. Le Breton. The Copeland measure of Condorcet choice functions. *Discrete Applied Mathematics*, 55:273–279, 1994.

J. Lang, M.-S. Pini, F. Rossi, K. B. Venable, and T. Walsh. Winner determination in sequential majority voting. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1372–1377, 2007.

J.-F. Laslier. *Tournament Solutions and Majority Voting*. Springer, 1997.

R. D. McKelvey and R. G. Niemi. A multistage game representation of sophisticated voting for binary procedures. *Journal of Economic Theory*, 18:1–22, 1978.

N. Miller. A new solution set for tournaments and majority voting: Further graph theoretical approaches to the theory of voting. *Americal Journal of Political Science*, 24:68–96, 1980.

H. Moulin. Choosing from a tournament. *Social Choice and Welfare*, 3:271–291, 1986.

A. D. Procaccia, A. Zohar, Y. Peleg, and J. S. Rosenschein. Learning voting trees. In *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI)*, pages 110–115, 2007.

R. Rivest and S. Vuillemin. On recognizing graph properties from adjacency matrices. *Theoretical Computer Science*, 3:371–384, 1976.

A. L. Rosenberg. The time required to recognize properties of graphs: A problem. *SIGACT News*, 5(4):15–16, 1973.

S. Srivastava and M. A. Trick. Sophisticated voting rules: The case of two tournaments. *Social Choice and Welfare*, 13:275–289, 1996.

A. C. Yao. Probabilistic computations: Towards a unified measure of complexity. In *Proceedings of the 17th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 222–227, 1977.

Felix Fischer
Institut für Informatik
Ludwig-Maximilians-Universität München
80538 München, Germany
Email: `fischerf@tcs.ifi.lmu.de`.

Ariel D. Procaccia and Alex Samorodnitsky
School of Computer Science and Engineering
The Hebrew University of Jerusalem
Jerusalem 91904, Israel
Email: `{arielpro,salex}@cs.huji.ac.il`.

From Preferences to Judgments and Back

Davide Grossi

Abstract

The paper studies the interrelationships of the Preference Aggregation and Judgment Aggregation problems from the point of view of logical semantics. The result of the paper is twofold. On the one hand, the Preference Aggregation problem is viewed as a special case of the Judgment Aggregation one. On the other hand, the Judgment Aggregation problem is viewed as a special case of the Preference Aggregation one. It is shown how to import an impossibility result from each framework to the other.

1 Introduction

Recent results [5, 9] have shown how Preference Aggregation theorems, such as Arrow's impossibility [1], can be obtained as corollaries of impossibility theorems concerning the aggregation of judgments. The idea behind this reduction consists in viewing preferences between issues as special kind of judgments, i.e., formulae to which a truth-value is attached. In [5, 9] the formulae used for representing preferences are first-order formulae of the type xPy ("x is strictly preferred to y") where x, y are variables for the elements in the set of issues of the Preference Aggregation problem. Then, in order for the judgments concerning such formulae to behave like a strict preference relation the three axioms of asymmetry, transitivity and connectedness[1] are added to the Judgment Aggregation framework.

The present paper proposes a different approach to obtain the same kind of reduction. More precisely, preferences will be studied as implicative statements $y \to x$ ("y is at most as preferred as x") in a many-value logic setting [6,7]. The insights gained by this reduction of preferences to judgments are then also used to obtain an inverse reduction of judgments to special kind of preferences, namely those preferences definable in the Boolean algebra on the support $\{0, 1\}$. To the best of our knowledge, this is the first work advancing a proposal on how to reduce Judgment Aggregation problems to Preference Aggregation ones.

The paper is structured as follows. In Section 2 the frameworks of Preference and Judgement Aggregation are briefly exposed, some basic terminology is introduced and some relevant results from the literature are summarized. In Section 3 a reduction of preferences to judgments is proposed which makes use of the semantics of many-valued logics. An impossibility result from Judgment Aggregation is thereby imported into Preference Aggregation. Section 4 explains how the framework of Preference Aggregation could be extended in order to incorporate preferences between complex issues representable as logical formulae. In Section 5 such idea is related to propositional logic and a reduction of judgments to preferences is proposed. Also in this case, an impossibility result of Preference Aggregation is imported into Judgment Aggregation. Section 6 briefly concludes.

2 Preliminaries

The present section is devoted to the introduction of the two frameworks of preference and judgment aggregation, and of some results which will be of use later in the paper.

[1] Asymmetry: $\forall x, y(xPy \to \neg yPx)$; Transitivity: $\forall x, y, z((xPy \wedge yPz) \to xPz)$; Connectedness: $\forall x, y(x \neq y \to (xPy \vee yPx))$

2.1 Preference Aggregation

Preference Aggregation (PA) concerns the aggregation of the preferences of several agents into one collective preference. A preference relation $\preceq$ on a set of issues $\mathtt{Iss}^P$ is a total preorder, i.e., a binary relation which is reflexive, transitive, and total. $\mathfrak{T}(\mathtt{Iss}^P)$ denotes the set of all total preorders of a set $\mathtt{Iss}^P$. As usual, on the ground of $\preceq$ we can define its asymmetric and symmetric parts: $x \prec y$ iff $(x,y) \in \preceq$ & $(y,x) \notin \preceq$; $x \approx y$ iff $(x,y) \in \preceq$ & $(y,x) \in \preceq$. We also define from $\preceq$ the following non-transitive relation: $x \lesseqgtr y$ iff $(x,y) \in \preceq$ or $(y,x) \in \preceq$ but not both. The notion of PA structure can now be defined.

Definition 1. *(Preference aggregation structure) A PA structure is a quadruple* $\mathfrak{S}^P = \langle \mathtt{Agn}^P, \mathtt{Iss}^P, \mathtt{Prf}^P, \mathtt{Agg}^P \rangle$ *where:* $\mathtt{Agn}^P$ *is a finite set of agents such that* $1 \leq |\mathtt{Agn}^P|$*;* $\mathtt{Iss}^P$ *is a finite set of issues such that* $3 \leq |\mathtt{Iss}^P|$*;* $\mathtt{Prf}^P$ *is the set of all preference profiles, i.e.,* $|\mathtt{Agn}^P|$*-tuples* $\mathfrak{p} = (\preceq_i)_{i \in \mathtt{Agn}^P}$ *where each* $\preceq_i$ *is a total preorder over* $\mathtt{Iss}$*;* $\mathtt{Agg}^P$ *is a function taking each* $\mathfrak{p} \in \mathtt{Prf}^P$ *to a total preorder over* $\mathtt{Iss}$*, i.e.,* $\mathtt{Agg}^P : \mathtt{Prf}^P \longrightarrow \mathfrak{T}(\mathtt{Iss}^P)$*. The value of* $\mathtt{Agg}^P$ *is denoted by* $\preceq$*.*

The aggregation function $\mathtt{Agg}^P$ is then studied under the assumption that it satisfies some of the following typical conditions:

Unanimity (**U**): $\forall x, y \in \mathtt{Iss}^P$ and $\forall \mathfrak{p} = (\preceq_i)_{i \in \mathtt{Agn}^P}$ if $\forall i \in \mathtt{Agn}^P, y \prec_i x$ then $y \prec x$.

Independence (**I**): $\forall x, y \in \mathtt{Iss}^P$ and $\forall \mathfrak{p} = (\preceq_i)_{i \in \mathtt{Agn}^P}, \mathfrak{p}' = (\preceq'_i)_{i \in \mathtt{Agn}^P}$, if $\forall i \in \mathtt{Agn}^P y \preceq_i x$ iff $y \preceq'_i x$, then $y \preceq x$ iff $y \preceq' x$.

Systematicity (**Sys**) $\forall x, y, x', y' \in \mathtt{Iss}^P$ and $\forall \mathfrak{p} = (\preceq_i)_{i \in \mathtt{Agn}^P}, \mathfrak{p}' = (\preceq'_i)_{i \in \mathtt{Agn}^P}$, if $\forall i \in \mathtt{Agn}^P y \preceq_i x$ iff $y' \preceq'_i x'$, then $y \preceq x$ iff $y' \preceq' x'$.

Non-dictatorship (**NoDict**): $\nexists i \in \mathtt{Agn}^P$ such that $\forall x, y \in \mathtt{Iss}^P$ and $\forall \mathfrak{p} = (\preceq_i)_{i \in \mathtt{Agn}^P}$, if $y \prec_i x$ then $y \prec x$.

Notice that Definition 1 incorporates also the aggregation conditions usually referred to as universal domain and collective rationality.

2.2 Judgment Aggregation

Judgment aggregation (JA) concerns the aggregation of sets of interrelated formulae into one collective set of formulae. This section introduces the framework for judgment aggregation based on the language of propositional logic [9].

A central notion in Judgment Aggregation is the notion of agenda. Intuitively, an agenda consists of all the possible positions that agents can assume about the truth and falsity of some issue.

Definition 2. *(JA agenda) The language of propositional logic is denoted by* $\mathcal{L}$*. Given a finite set of formulae* $\mathtt{Iss}^J \subseteq \mathcal{L}$*, the set* $\mathfrak{ag}(\mathtt{Iss}^J) = \{\models \phi \mid \phi \in \mathtt{Iss}^J\} \cup \{\not\models \phi \mid \phi \in \mathtt{Iss}^J\}$ *denotes the* agenda *of* $\mathtt{Iss}^J$*.*

A JA agenda consists therefore of a set of pairs $(\models \phi, \not\models \phi)$, each member in the pairs meaning that ϕ is assigned value 1 and, respectively, a different value from 1, which is in the propositional case 0. It might be worth noticing that we assume a slightly different perspective on agendas than the literature on JA. Normally, an agenda is viewed as a set of position/negation pairs $(\phi, \neg\phi)$. In this view, the judgment themselves can be seen as formulae of the language from which the issues are drawn. Instead, we prefer to see judgments as meta-formulae stating whether a given issue is accepted (true) or rejected (not true). In propositional logic, however, the two perspectives are equivalent.

A judgment set picks one element out of each such pairs keeping propositional consistency. More formally, a *judgment set* for the agenda $\mathfrak{ag}(\mathtt{Iss}^J)$ is a set $J \subseteq \mathfrak{ag}(\mathtt{Iss}^J)$ which is consistent and complete ($\forall \phi \in \mathtt{Iss}^J$: either $\models \phi \in J$ or $\not\models \phi \in J$ but not both) and closed under propositional logic consequence ($\forall \phi, \psi \in \mathtt{Iss}^J$: if $\phi \models \psi$ and $\models \phi \in J$ then $\models \psi \in J$). The set of all judgment sets for the agenda built on $\mathtt{Iss}^J$ is denoted by $\mathfrak{J}(\mathtt{Iss}^J)$. The set $\mathtt{Iss}_0^J$ denotes the set of propositional atoms in $\mathtt{Iss}^J$.

Definition 3. *(Judgment aggregation structure) A judgment aggregation structure is a quadruple* $\mathfrak{S}^J = \langle \mathit{Agn}^J, \mathit{Iss}^J, \mathit{Prf}^J, \mathit{Agg}^J \rangle$ *where:* Agn^J *is a finite set of agents such that* $1 \leq |\mathit{Agn}^J|$*;* Iss^J *is a finite set of issues consisting of propositional formulae and containing at least two atoms:* $\mathit{Iss}^J \subseteq \mathcal{L}$ *s.t.* $2 \leq |\mathit{Iss}_0^J|$*;* Prf^J *is the set of all judgment profiles, i.e.,* $|\mathit{Agn}^J|$*-tuples* $\mathtt{j} = \{J_i\}_{i \in \mathit{Agn}^J}$ *where each* J_i *is a judgment set for the agenda* $\mathfrak{ja}(\mathit{Iss}^J)$*;* Agg^J *is a function taking each* $\mathtt{j} \in \mathit{Prf}^J$ *to a judgment set for* $\mathfrak{ag}(\mathit{Iss}^J)$*, i.e.,* $\mathit{Agg}^J : \mathit{Prf}^J \longrightarrow \mathfrak{J}(\mathit{Iss}^J)$*. The value of* Agg^J *is denoted by* J*.*

Like in PA, in JA aggregation functions are studied under specific conditions. The following conditions reformulate the ones proper of PA:

Unanimity (**U**): $\forall x \in \mathtt{Iss}^J$ and $\forall \mathtt{j} = (J_i)_{i \in \mathtt{Agn}^J}$ if $\forall i \in \mathtt{Agn}^J, \models x \in J_i$ then $\models x \in J$ and if $\not\models x \in J_i$ then $\not\models x \in J$.

Independence (**I**): $\forall x \in \mathtt{Iss}^J$ and $\forall \mathtt{j} = (J_i)_{i \in \mathtt{Agn}^J}, \mathtt{j}' = (J_i')_{i \in \mathtt{Agn}^J}$, if $\forall i \in \mathtt{Agn}^J \models x \in J_i$ iff $\models x \in J_i'$ then $\models x \in J$ iff $\models x \in J'$.

Systematicity (**Sys**): $\forall x, y \in \mathtt{Iss}^J$ and $\forall \mathtt{j} = (J_i)_{i \in \mathtt{Agn}^J}, \mathtt{j}' = (J_i')_{i \in \mathtt{Agn}^J}$, if $\forall i \in \mathtt{Agn}^J \models x \in J_i$ iff $\models y \in J_i'$, then $\models x \in J$ iff $\models y \in J'$, and if $\models x \in J_i$ iff $\not\models y \in J_i'$ then $\models x \in J$ iff $\not\models y \in J'$.

Non-dictatorship (**NoDict**): $\nexists i \in \mathtt{Agn}^J$ such that $\forall x \in \mathtt{Iss}^J$ and $\forall \mathtt{j} = (J_i)_{i \in \mathtt{Agn}^J}$, if $\models x \in J_i$ then $\models x \in J$ and if $\not\models x \in J_i$ then $\not\models x \in J$.

Notice that Definition 3 incorporates the aggregation conditions usually referred to as universal domain and collective rationality. In the remainder of the paper we will refer to the conditions of unanimity, independence, systematicity and non-dictatorship as **U**, **I**, **Sys**, respectively, **NoDict**. It will be clear from the context whether the condition at issue should be interpreted in its PA or in its JA formulation.

2.3 Some relevant results on JA and PA

We now briefly sketch two results which are of importance for the development of the work presented in this paper: Propositions 1 and 2.

JA agendas can be studied from the point of view of their structural properties, that is, how strictly connected are the issues with which the agenda is concerned. In [5] the following structural property —among others— is studied.

Definition 4. *(Minimal connectedness of the agenda) An agenda* $\mathfrak{ag}(\mathit{Iss}^J)$ *is minimally connected if: i) it has a minimal inconsistent subset* $S \subseteq \mathfrak{ag}(\mathit{Iss}^J)$ *such that* $3 \leq |S|$*; ii) it has a minimal inconsistent subset* $S \subseteq \mathfrak{ag}(\mathit{Iss}^J)$ *such that* $(S \backslash Z) \cup \{\not\models x \mid \ \models x \in Z\}$ *is consistent for some* $Z \subseteq \mathfrak{ag}(\mathit{Iss}^J)$ *of even size.*

A typical JA impossibility result making use of this property is the following [5, 10]. In Section 3 this result will then be imported, as an example, from JA to PA.

Proposition 1. *(Impossibility for minimally connected agendas) For any JA structure* $\mathfrak{S}^J$ *there exists no aggregation function for a minimally connected agenda* $\mathfrak{ag}(\mathit{Iss}^J)$ *which satisfies* **U***,* **Sys** *and* **NoDict***.*

$\{x,y\}$	$\{y,z\}$	$\{x,z\}$	$\{x,y\}$	$\{y,z\}$	$\{x,z\}$
$y \prec x$	$z \prec y$	$z \prec x$	$u(y) < u(x)$	$u(z) < u(y)$	$u(z) < u(x)$
$y \prec x$	$y \prec z$	$x \prec z$	$u(y) < u(x)$	$u(y) < u(z)$	$u(x) < u(z)$
$x \prec y$	$z \prec y$	$x \prec z$	$u(x) < u(y)$	$u(z) < u(y)$	$u(x) < u(z)$
$y \prec x$	$z \prec y$	$x \prec z$	$u(y) < u(x)$	$u(z) < u(y)$	$u(x) < u(z)$

Table 1: Condorcet's paradox.

Proof. We refer the reader to [5]. □

A ranking function attributes a ranking (or value, or utility, or payoff) to all the issues of a preference aggregation problem. In this paper we will make use of ranking functions with the real interval $[0, 1]$ as codomain. For such functions the following result holds which is the special case of a theorem first proven in [4]. This result will play a central role in the next section.

Proposition 2. *(Representation of $\preceq$) Let $\preceq$ be a total preorder on a finite set X. There exists a ranking function $u : X \longrightarrow [0, 1]$ such that $\forall x, y \in X$: $x \preceq y$ iff $u(x) \leq u(y)$. Such a function is unique up to ordinal transformations*[2]*.*

Proof. The reader is referred to [4]. □

3 Preferences as judgments

This section is devoted to show how the aggregation of preferences can be studied in terms of the aggregation of judgments. As anticipated in Section 1 we get to the very same conclusions presented in [5]. Nevertheless, to obtain such result we will follow a different approach based on logical semantics rather than logical axiomatics. Such approach will offer, in Section 5, also a method for viewing judgments as forms of preferences.

3.1 Condorcet's paradox as a judgment aggregation paradox

In Condorcet's paradox, pairwise majority voting on issues generates a collective preference which is not transitive. From Proposition 2 we know that any preference relation which is a total preorder can be represented by an appropriate ranking function u with codomain $[0, 1]$. Table 1 depicts the standard version of the paradox in relational notation, and the version using ranking functions u. To obtain the paradox strict preferences are not necessary. The paradox arises also in weaker forms like the one depicted in Table 2. Again, both the relational and the ranking function-based versions are provided.

The basic intuition underlying this section consists in reading the right-hand sides of Tables 1 and 2 as if u was an interpretation function of propositions x, y, z on the real interval $[0, 1]$. In many-valued logic [6, 7], a semantic clause such as $u(y) \leq u(x)$ typically defines the satisfaction by u of the implication $y \rightarrow x$:

$$u \models y \rightarrow x \text{ iff } u(y) \leq u(x). \tag{1}$$

Intuitively, implication $y \rightarrow x$ is true (or accepted, or satisfied) iff the rank of y is at most as high as the rank of x. Since we know by Proposition 2 that any total preorder $\preceq$ can

[2]We recall that an ordinal transformation t is a function such that for all utilities m and n, $t(m) \leq t(n)$ iff $m \leq n$.

$\{x,y\}$	$\{y,z\}$	$\{x,z\}$
$y \preceq x$	$z \preceq y$	$z \preceq x$
$y \preceq x$	$y \prec z$	$x \prec z$
$x \prec y$	$z \preceq y$	$x \prec z$
$y \preceq x$	$z \preceq y$	$x \prec z$

$\{x,y\}$	$\{y,z\}$	$\{x,z\}$
$u(y) \leq u(x)$	$u(z) \leq u(y)$	$u(z) \leq u(x)$
$u(y) \leq u(x)$	$u(y) < u(z)$	$u(x) < u(z)$
$u(x) < u(y)$	$u(z) \leq u(y)$	$u(x) < u(z)$
$u(y) \leq u(x)$	$u(z) \leq u(y)$	$u(x) < u(z)$

Table 2: Weak Condorcet's paradox.

be represented by a ranking function, the bridge between the notion of preference and an equivalent notion of judgment is thereby readily available: given a total preorder $\preceq$, there always exists a ranking function u, unique up to order-preserving transformations such that: $y \preceq x$ iff $u(y) \leq u(x)$. We thus obtain a direct bridge between preferences and judgments. In fact, by exploiting Formula 1 the right-hand side of Table 2 can be rewritten as in Table 3. Notice that x is substituted by p, y by q and z by r. The same could obviously be done for Table 1.

In other words, by first reading the weak Condorcet's paradox in terms of ranking function (Proposition 2), and then interpreting such functions from the point of view of logical semantics (Formula 1), we can show the equivalence between a preference aggregation problem and a judgment aggregation one. The following result generalizes this observation.

Proposition 3. *(From preferences to judgments) The set of all PA structures can be mapped into the set of all JA structures in such a way that each of them corresponds exactly to one JA structure whose set of issues $\mathtt{Iss}^J$ consists of only implications.*

Proof. We show how to construct a structure $\mathfrak{S}^J$ from any structure $\mathfrak{S}^P$. Let $\mathfrak{S}^P = \langle \mathtt{Agn}^P, \mathtt{Iss}^P, \mathtt{Prf}^P, \mathtt{Agg}^P \rangle$. The set of agents $\mathtt{Agn}^J$ of $\mathfrak{S}^J$ is the same: $\mathtt{Agn}^P = \mathtt{Agn}^J$. The set of issues $\mathtt{Iss}^J$ is such that: $\mathtt{Iss}_0^J = \mathtt{Iss}^P$, that is, $\mathtt{Iss}^P$ provides the propositional atoms of $\mathtt{Iss}^J$; and it contains all implications built from $\mathtt{Iss}_0^J$. The set $\mathtt{Prf}^J$ is the set of all judgment profiles obtained by translating any total order $\preceq_i$ into a judgment set J_i as follows: $\models x \to y \in J_i$ iff $(x,y) \in \preceq_i$. Notice that, as a consequence, we can define a bijection $\mathtt{bi}$ between $\mathtt{Prf}^P$ and $\mathtt{Prf}^J$. Finally, the aggregation function $\mathtt{Agg}^J$ can be defined as follows: $\mathtt{Agg}^J(\mathtt{bi}(\mathfrak{p})) = \mathtt{bi}(\mathtt{Agg}^P(\mathfrak{p}))$. This completes the construction. □

The JA structures $\mathfrak{S}^J$ resulting from the construction in the proof are such that their set of issues $\mathtt{Iss}^J$ consist of all implications obtained from a set of atoms $\mathtt{Iss}_0^J$ such that $3 \leq |\mathtt{Iss}_0^J|$. We call an agenda $\mathfrak{ag}(\mathtt{Iss}^J)$ built on such a set $\mathtt{Iss}^J$ an *implicative agenda*.

$\{p,q\}$	$\{q,r\}$	$\{p,r\}$
$\models q \to p$	$\models r \to q$	$\models r \to p$
$\models q \to p$	$\not\models r \to q$	$\not\models r \to p$
$\not\models q \to p$	$\models r \to q$	$\not\models r \to p$
$\models q \to p$	$\models r \to q$	$\models r \to p$

$q \to p$	$r \to q$	$r \to p$
$\models$	$\models$	$\models$
$\models$	$\not\models$	$\not\models$
$\not\models$	$\models$	$\not\models$
$\models$	$\models$	$\not\models$

Table 3: Weak Condorcet's paradox as a judgment aggregation paradox.

3.2 Doing PA in JA

We know how to translate a PA structure into a JA one with implicative agendas. As an example of how to import impossibility results of JA to PA we apply Proposition 1 to JA structures with implicative agendas. In order to do so, we first need to show that implicative agendas enjoy the property of minimal connectedness.

Proposition 4. *Implicative agendas are minimally connected.*

Proof. Suppose $\{p, q, r\} \subseteq \mathtt{Iss}_0^J$. We prove that the implicative agenda built on $\mathtt{Iss}_0^J$ is minimally connected. The desired minimal inconsistent subset of the agenda with size higher than or equal to 3, and such that by negating two of its judgments consistency is restored, is: $\{\models q \to p, \models r \to q, \not\models r \to p\}$. □

Everything is in place now to prove an impossibility result concerning the set of JA structures in which PA can be mapped, i.e., those with implicative agendas.

Theorem 1. *(A JA theorem for PA) For any PA structure $\mathfrak{S}^P$, there exists no aggregation function which satisfies* **U***,* **Sys** *and* **NoDict**.

Proof. The theorem follows from Propositions 1, 3 and 4. □

The theorem itself is, of course, not surprising. It is just an illustration of the embedding advanced in this section. More interesting results can be obtained along the very same lines by studying different structural properties of the agenda, e.g., strong connectedness [5].

Before closing the present section it is worth spending a few words on the relation between the approach presented here and the one presented in [5] where Arrow's theorem is proven as a corollary of a JA theorem analogous to Proposition 1 and a "bridging" proposition analogous to Proposition 4. Both approaches somehow reduce PA to JA, but while [5] does it axiomatically by imposing further constraints on a first-order logic agenda (i.e., the axioms of strict total orders), we do it semantically, by ranking the issues in $\mathtt{Iss}^P$ on the $[0, 1]$ interval and interpreting preferences as implications. A more in-depth comparison of the two approaches deserves further investigation. However, the semantic view has the advantage of hinting also a "way back" from judgments to preferences. Such "way back" is the topic of the next two sections.

4 Intermezzo: Rankings as Truth Values

The essential difference between JA and PA is that, in JA, issues display logical form. Is there a consistent way to talk about compound issues in PA obtained by performing logical operations on atomic ones? In other words, is there a way to define preferences which display the logical complexity of judgments? Aim of the present section is to shows how to answer these questions by generalizing Formula 1 to any logical connective.

Once we consider the set of issues $\mathtt{Iss}^P$ of a preference aggregation problem $\mathfrak{S}^P$ to be the finite set of propositional atoms $\mathcal{L}_0$ of a propositional language $\mathcal{L}$, any ranking function u can be viewed as an interpretation function of the atoms in $\mathcal{L}_0$ on the real interval $[0, 1]$. The natural question follows: how to inductively extend a function u in order to interpret issues in $\mathtt{Iss}^P$ which consist of propositional formulae, and not just atoms? This question takes us into the realm of many-valued logics, and many possibilities are available. However, given Formula 1, we are looking for something in particular. We want an implication to be satisfied exactly when the antecedent is ranked at most as high as the consequent. More precisely, let us denote with $\breve{u}$ the inductive extension of the ranking function u. What we

are looking for is a multi-valued logic such that the following holds for any ranking function u and formulae ϕ, ψ:

$$u \models \phi \rightarrow \psi \quad \text{iff} \quad \breve{u}(\phi) \leq \breve{u}(\psi) \quad \text{iff} \quad \breve{u}(\phi \rightarrow \psi) = 1. \tag{2}$$

That is to say, the desired logic should be able to encode in the language the total order $\leq$ of rankings, so that $\breve{u}$ assigns the maximum ranking 1 to $\phi \rightarrow \psi$ (i.e., $\phi \rightarrow \psi$ is satisfied by u) iff the value assigned by $\breve{u}$ to ϕ is at most the same value assigned by $\breve{u}$ to ψ. The intuition behind Formula 2 consists in viewing the maximum ranking as the designated value for expressing the truth of compound formulae, and in particular of those formulae which express preferences between other formulae.

The property expressed in Formula 2 turns out to be a typical property of the family of t-norm multi-valued logics, or logics based on triangular norms [7]. In such logics the connective $\rightarrow$ denotes the algebraic residuum operation on truth degrees (i.e., rankings). Residua come always in pairs with t-norm operations, so what is going to characterize the logic we are looking for is the t-norm we choose to be paired with the residuum denoted by $\rightarrow$. The most straightforward candidate is the algebraic infimum, denoted by the standard logic conjunction $\wedge$. To sum up, we want that the following holds for any ranking function u and formulae ϕ, ψ, ξ:

$$\breve{u}(\phi \wedge \xi) \leq \breve{u}(\psi) \quad \text{iff} \quad \breve{u}(\xi) \leq \breve{u}(\phi \rightarrow \psi). \tag{3}$$

If we then take the rest of the connectives $\neg$ and $\vee$ to denote, as usual, the algebraic complementation and, respectively, the algebraic supremum, the many-valued logic satisfying Formulae 2 and 3 is the logic known as Gödel-Dummett logic (**GD** in short).

By using the semantics of **GD** (i.e., Gödel algebra) it is possible to extend the PA framework in order to incorporate preferences between compound issues represented as logical formulae. What kind of new insights this extension provides in the study of the aggregation of preferences deserves further research, but it falls outside the scope of the present study. In fact, what we are interested in now is to show that the insights provided by **GD** on the representation of preferences between logical formulae can be used, after a slight modification, in order to view judgments as preferences, thereby providing a reduction of the JA problem to the PA problem. To this aim is devoted the next section.

5 Judgments as Preferences

In its original form, JA is based on propositional logic. The considerations of the last section about how to interpret compound issues in a PA structure can be directly used for representing judgments as preferences. The key step consists in considering that propositional logic is the extension of **GD** with the bivalence principle. Like an interpretation function of **GD** (i.e., a ranking function u) determines a total preorder on the set of issues of a PA structure, so does a propositional interpretation function on the set of issues of a JA structure. Obviously, the type of total preorder yielded by a propositional interpretation function is of a specific kind.

5.1 Boolean preference profiles

Like **GD** is sound and complete w.r.t. Gödel algebra [7], propositional logic is sound and complete w.r.t. $\mathbf{2} = \langle \{0,1\}, \sqcap, \sqcup, -, 0, 1 \rangle$, i.e., the Boolean algebra on the support $\{1, 0\}$, where $\sqcap$ and $\sqcup$ are the `min`, respectively, `max` operations, $-$ is the involution defined as $-x = 1 - x$, and 0 and 1 are the designated elements. The total preorders generated by a propositional interpretation function are called *Boolean preferences*.

Definition 5. *(Boolean preferences) A Boolean preference is a total preorder on a set of formulae Φ closed under atoms which can be mapped to $\langle\{1,0\},\leq\rangle$ by a function $\breve{v}:\Phi\longrightarrow\{1,0\}$ such that: $\forall\phi,\psi\in \mathit{Iss}^J$, $\phi\preceq\psi$ iff $\breve{v}(\phi)\leq\breve{v}(\psi)$; and $\breve{v}$ preserves the meaning of propositional connectives, that is:*

$$\breve{v}(\top)=1,\quad \breve{v}(\neg\phi)=1-\breve{v}(\phi),\quad \breve{v}(\phi\wedge\psi)=\mathtt{min}(\breve{v}(\phi),\breve{v}(\psi)),\quad \breve{v}(\phi\vee\psi)=\mathtt{max}(\breve{v}(\phi),\breve{v}(\psi)).$$

A few considerations are in order. Notice that, within a Boolean preference, $\prec$-paths have maximum length 1. In fact, stating that $y\prec x$ is equivalent to assign value 1 to x and value 0 to y. Notice also that a total preorder containing $x\prec y, y\prec x\wedge y$ cannot be a Boolean preference since there exists no function assigning 1 and 0 to x and y, which preserves $\preceq$ on $\leq$ and, at the same time, $\wedge$ on $\mathtt{min}$. Notice also that Boolean preferences generalize dichotomous preferences[3] allowing issues to be compounded following the standard algebraic semantics of logical connectives, and allowing issues to be all ranked as maximal or minimal. In fact, dichotomous preferences are nothing but Boolean preferences over atoms, i.e., logically unrelated issues, s.t. the sets of maximal and minimal elements always contain at least one atom (i.e., the preferences are never "unconcerned" [3]). In addition, Boolean preferences feature the constants $\top$ and $\bot$. These denote trivial issues which every voter places in the sets of $\preceq$-maximal and, respectively, $\preceq$-minimal elements. The following holds.

Proposition 5. *(From judgment sets to Boolean preferences) Every judgment set J on the set of issues Iss^J can be translated to a Boolean preference $\preceq$ over Iss^J such that $\models\phi\in J$ iff $\phi\approx\top$.*

Proof. Since each judgment set J is closed under atoms, it univocally determines a propositional evaluation $\breve{v}:\mathtt{Iss}^J\longrightarrow\{1,0\}$. The evaluation is an homomorphism from the formula algebra built on the set of atoms in $\mathtt{Iss}^J$ and $\mathbf{2}$. In $\mathbf{2}$ the partial order $\leq$ can be defined in the usual way: $x\leq y:=\mathtt{min}(x,y)=x$. It is easy to see that $\leq$ is then a total preorder. Function $\breve{v}$ of J can therefore be used to univocally define a total preorder $\preceq$ on $\mathtt{Iss}^J$ as follows: $\forall\phi,\psi\in\mathtt{Iss}^J$, $\phi\preceq\psi$ iff $\breve{v}(\phi)\leq\breve{v}(\psi)$. It follows that $\preceq$ is a Boolean preference by construction. From left to right. If $\models\phi\in J$ then $\breve{v}(\phi)=1$, therefore, by construction $\phi\approx\top$. From right to left. If $\phi\approx\top$ then $\breve{v}(\phi)=\breve{v}(\top)$, hence $\breve{v}(\phi)=1$ □

Intuitively, Proposition 5 guarantees the analogue of Proposition 3 to hold. In other words, every JA structure $\mathfrak{S}^J$ can be translated to an equivalent PA structure[4] by just stating $\mathtt{Iss}^P=\mathtt{Iss}^J$ and letting $\mathtt{Prf}^P$ be the set of Boolean preference profiles on $\mathtt{Iss}^P$. We thus obtain a way to view judgments as preferences. As an example, in the next section, we will consider the Discursive paradox from a PA point of view.

5.2 The Discursive paradox as a PA paradox

In the Discursive paradox propositionwise majority voting leads to the definition of an impossible evaluation of the propositions at issue, as showed in the left-hand side of Table 4. The middle of Table 4 depicts the Discursive paradox by means of the two truth-values. Finally, if we consider these truth values as rankings of the propositions at issue, we get to the right-hand side of the table. There a PA version of the paradox is depicted. Recall that $x\approx y$ iff $(x,y)\in\preceq$ & $(y,x)\in\preceq$, and $x\lessapprox y$ iff $(x,y)\in\preceq$ or $(y,x)\in\preceq$ but not both. In this case the aggregated preference violates the transitivity of $\approx$.

As such, the Discursive paradox can fruitfully be viewed as yet another variant of Condorcet's paradox.[5] In fact, in Condorcet's paradox the collective preference violates the

[3]Dichotomous preferences are such that $\forall x,y,z\in\mathtt{Iss}^P$, either $x\approx y$ or $y\approx z$ or $x\approx z$ but not all [8].

[4]Notice that the same does not hold for dichotomous preferences where we cannot distinguish between preferences ranking all issues equal to $\top$ or all equal to $\bot$ (see also Proposition 7).

[5]Other representations are possible by making use of the constants $\top$ or $\bot$, and indifference $\approx$.

p	$p \to q$	q
$\models$	$\models$	$\models$
$\models$	$\not\models$	$\not\models$
$\not\models$	$\models$	$\not\models$
$\models$	$\models$	$\not\models$

p	$p \to q$	q
1	1	1
1	0	0
0	1	0
1	1	0

$\{p, p \to q\}$	$\{p, q\}$	$\{q, p \to q\}$
$p \approx p \to q$	$p \approx q$	$q \approx p \to q$
$p \lessapprox p \to q$	$q \lessapprox p$	$q \approx p \to q$
$p \lessapprox p \to q$	$p \approx q$	$q \lessapprox p \to q$
$p \lessapprox p \to q$	$p \approx q$	$q \approx p \to q$

Table 4: Discursive paradox as a PA paradox.

transitivity of $\prec$, in the weak Condorcet analyzed in Section 3 what is violated is instead the transitivity of $\preceq$, here it is the transitivity of $\approx$.

5.3 Arrow's conditions and Boolean preferences

We can now represent judgments as special kinds of preferences. Can we also import PA impossibility results to JA? A first natural step in this direction is to study how Boolean preferences behave under the standard Arrow's conditions: **U**, **I** and **NoDict**. Under these conditions, and assuming the domain and codomain of the aggregation function to be drawn from the set of Boolean preferences, yields an impossibility. However, the source of the impossibility does not rest upon the logical connectives, as is typically the case in JA.

Proposition 6. *(Arrow's theorem holds for dichotomous domains and codomains) Let $\mathfrak{S}^P$ contain a set of issues $\mathtt{Iss}^P$ s.t. $3 \leq |\mathtt{Iss}^P|$, and $\mathtt{Prf}^P$ is the set of dichotomous preference profiles on $\mathtt{Iss}^P$. There exists no aggregation function which satisfies* **U***,* **I** *and* **NoDict***.*

Proof. See Appendix. □

Limiting the domain and codomain of the aggregation function to dichotomous preferences does not resolve the impossibility. Since dichotomous preferences are a subset of Boolean preferences the result carries over. Conditions **U**, **I** and **NoDict** seem therefore to be too strong to yield JA-like impossibilities. This is not surprising because, unlike in the translation from PA to JA (Section 3), the JA formulation of the conditions cannot be directly obtained from their PA formulation. The translation of the standard JA conditions in a Boolean preferences format look instead like this:

($\mathbf{U}^\top$): $\forall x \in \mathtt{Iss}^P$ and $\forall \mathfrak{p} = (\preceq_i)_{i \in \mathtt{Agn}^P}$ if $\forall i \in \mathtt{Agn}^P, x \approx_i \top$ then $x \approx \top$ and if $x \approx_i \bot$ then $x \approx \bot$.

($\mathbf{I}^\top$): $\forall x \in \mathtt{Iss}^P$ and $\forall \mathfrak{p} = (\preceq_i)_{i \in \mathtt{Agn}^P}, \mathfrak{p}' = (\preceq'_i)_{i \in \mathtt{Agn}^P}$, if $\forall i \in \mathtt{Agn}^P x \approx_i \top$ iff $x \approx'_i \top$ then $x \approx \top$ iff $x \approx' \top$.

($\mathbf{Sys}^\top$): $\forall x, y \in \mathtt{Iss}^P$ and $\forall \mathfrak{p} = (\preceq_i)_{i \in \mathtt{Agn}^P}, \mathfrak{p}' = (\preceq'_i)_{i \in \mathtt{Agn}^P}$, if $\forall i \in \mathtt{Agn}^J x \approx_i \top$ iff $y \approx'_i \top$, then $x \approx \top$ iff $y \approx' \top$, and if $x \approx_i \top$ iff $y \approx'_i \bot$, then $x \approx \top$ iff $y \approx' \bot$.

($\mathbf{NoDict}^\top$): $\nexists i \in \mathtt{Agn}^P$ such that $\forall x \in \mathtt{Iss}^P$ and $\forall \mathfrak{p} = (\preceq_i)_{i \in \mathtt{Agn}^P}$, if $x \approx_i \top$ then $x \approx \top$ and if $x \approx_i \bot$ then $x \approx \bot$.

It is straightforward to see that these conditions exactly correspond to the standard JA conditions exposed in Section 2. The following proposition compares the relative strength of these JA-like formulations w.r.t. the standard PA-like ones.

Proposition 7. *(JA vs. PA conditions) The following relations hold under Boolean preferences:* $\mathbf{U}^\top$ *implies* **U***;* $\mathbf{I}^\top$ *implies* **I***;* $\mathbf{Sys}^\top$ *implies* **Sys***;* **NoDict** *implies* $\mathbf{NoDict}^\top$*. The converses do not hold.*

Proof. The direction from left to right is trivial. The failure of the converses is due to the fact that for profiles in which all agents are indifferent w.r.t. all issues, it cannot be inferred whether they are all ranked equal to $\bot$ or equal to $\top$. □

Notice that in the case of non-dictatorship the implication has a contrapositive form. In fact, **NoDict** is stronger than $\mathbf{NoDict}^{\top}$. That is, the non-existence of a dictator in the standard Arrowian sense implies the non-existence of a dictator in the JA sense. This observation suggests that results such as Proposition 6 set out to prove impossibilities on the ground of a notion of non-dictatorship which is stronger than the one typically used in JA. The next section fine-tunes Proposition 6 obtaining an appropriate JA impossibility in the framework of Boolean preferences.

5.4 Doing JA in PA

In order to obtain an impossibility for Boolean preferences under $\mathbf{NoDict}^{\top}$, the conditions $\mathbf{U}^{\top}$ and $\mathbf{I}^{\top}$ do not suffice and, perhaps not surprisingly, $\mathbf{I}^{\top}$ needs to be substituted by $\mathbf{Sys}^{\top}$. We are thus in the position to prove a JA theorem within PA.

Theorem 2. *(Impossibility for Boolean Preferences) Let $\mathfrak{S}^P$ contain a set of issues $\mathtt{Iss}^P$ s.t. $\{p, q, p \wedge q\}$ (where $\wedge$ can be substituted by $\vee$ or $\rightarrow$) and $\mathtt{Prf}^P$ is the set of Boolean preference profiles on $\mathtt{Iss}^P$. There exists no aggregation function which satisfies $\mathbf{U}^{\top}$, $\mathbf{Sys}^{\top}$ and $\mathbf{NoDict}^{\top}$.*

Proof. See Appendix. □

Now, Proposition 5 guarantees that each JA structure $\mathfrak{S}^J$ can be translated to an equivalent PA structure $\mathfrak{S}^P$ over Boolean preferences, and the standard JA aggregation conditions can be directly translated to conditions conditions $\mathbf{U}^{\top}$, $\mathbf{Sys}^{\top}$ and $\mathbf{NoDict}^{\top}$. It follows that Theorem 2 provides an impossibility result for the aggregation of judgments. Notice also that such result is reminiscent of several JA theorems available in the literature (e.g. in [9]). To conclude, the basic argument of the section runs as follows: judgment sets univocally determine Boolean preferences, a peculiar form of Arrow's impossibility holds also for Boolean preference domains, hence it can be imported into JA.

6 Conclusions

By borrowing ideas from logical semantics, the paper has shown that, on the one hand, PA can be viewed as a special case of JA [5, 9] and that, on the other hand, the converse also holds. This suggests that PA and JA could be studied as the two faces of a same coin. It is our claim that the study of such interrelationship can be fruitfully pushed further by cross importing more (im)possibility results, which we plan to do in future researches.

Acknowledgments

The author is very grateful to Gabriella Pigozzi, Leon van der Torre, Paul Harrenstein and Franz Dietrich. Their comments have greatly improved the present version of the paper.

References

[1] K. Arrow. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4):328–346, 1950.

[2] K. Arrow. *Social Choice and Individual Values*. John Wiley, New York, 1963.

[3] S. J. Brams and P. Fishburn. Approval voting. *American Political Science Review*, 72:831–847, 1978.

[4] G. Debreu. Represetation of a preference ordering by a numerical function. In R. M. Thrall, C. H. Coombs, and R. L. Davis, editors, *Decision Processes*. John Wiley, 1954.

[5] F. Dietrich and C. List. Arrow's theorem in judgment aggregation. *Social Choice and Welfare*, 29(19–33), 2007.

[6] S. Gottwald. Many-valued logics. In D. Jacquette, editor, *Handbook of the Philosophy of Sciences*, volume 5. North-Holland, 2007.

[7] R. Hähnle. Advanced many-valued logics. In D. M. Gabbay and F. Guenther, editors, *Handbook of Philosophical Logic, 2nd Edition*, volume 2, pages 297–395. Kluwer, 2001.

[8] K. Inada. A note on the simple majority decision rule. *Econometrica*, 32(4):525–531, 1964.

[9] C. List and P. Pettit. Aggregating sets of judgments: Two impossibility results compared. *Synthese*, 140:207–235, 2004.

[10] M. Pauly and M. van Hees. Logical constraints on judgment aggregation. *Journal of Philosophical Logic*, 35:569–585, 2006.

A Proofs of Proposition 6 and Theorem 2

A proof can be obtained along the same lines of Arrow's original proof [2]. Let us first introduce some terminology. A set $V \subseteq \mathtt{Agn}^P$ is *almost decisive* for issue x over issue y (in symbols, $AD_V(x,y)$) iff: if $\forall i \in V, y \prec_i x$ and $\forall i \notin V, x \prec_i y$ then $y \prec x$. A set $V \subseteq \mathtt{Agn}^P$ is *decisive* for issue x over issue y (in symbols, $D_V(x,y)$) iff: if $\forall i \in V, y \prec_i x$ then $y \prec x$. Obviously, for any $x, y \in \mathtt{Iss}^P$: $D_V(x,y)$ implies $AD_V(x,y)$. We first need two following lemmata.

Lemma 1. *Let $\mathfrak{S}^P$ be such that $\mathtt{Iss}^P$ is a set of issues s.t. $3 \leq |\mathtt{Iss}^P|$, and $\mathtt{Prf}^P$ is the set of dichotomous preferences on $\mathtt{Iss}^P$. If there exists an individual $i \in \mathtt{Agn}^P$ such that $AD_V(x,y)$ for some pair (x,y) then, under the conditions* **U** *and* **I**, *i is decisive for any pair of issues, that is, i is a dictator.*

Proof. There are 6 pairs of issues with respect to which i can be almost decisive for the first element in the pair over the second one. For each of these pairs (x,y) we show that if $AD_i(x,y)$ then i is decisive for at least one of the remaining pairs $(x,z),(z,x),(y,z),(z,y)$ where z is the third issue in $\mathtt{Iss}^P$. Let I denote $\mathtt{Agn}^P - \{i\}$. $[AD_i(x,y) \Rightarrow D(x,z)]$ Assume $AD_i(x,y)$ and suppose the following holds: $y \prec_i x, z \approx_i y, z \prec_i x$ and $x \prec_I y$. In the collective preference, which is dichotomous, one of the following must hold: A) $y \prec x, y \prec z$; or B) $y \prec x, z \approx y$ However, A) violates **U** and **I**. To prove this latter claim, consider a profile which is identical to the given one w.r.t. the preferences of i and I over y, z, but such that $x \prec_i y$ in the preference profile we assumed. By **U**, in the collective preference for this new profile $x \prec y$ holds. By **I**, $y \prec z$ should also hold, which is impossible since the collective preference should be dichotomous. Therefore, B) is the only possible collective preference. Since B) is dichotomous, it holds that $z \prec x$, hence $AD_i(x,y)$ implies $D(x,z)$. We can argue symmetrically to prove that $AD_i(x,y) \Rightarrow D(z,y)$. Hence, $AD_i(x,y)$ implies $D(x,z)$ and $D(z,y)$. It is then sufficient to argue via permutations of the possible alternatives,

e.g., by interchanging y with z we obtain that $AD_i(x,z)$ implies $D(x,y)$ and $D(y,z)$. As a consequence, if $AD_i(x,y)$ agent i results to be decisive for any pair of alternatives drawn from the set $\{x,y,z\}$. This completes the proof of the lemma. □

Lemma 2. *Let $\mathfrak{S}^P$ be such that $\mathtt{Iss}^P$ is a set of issues s.t. $3 \leq |\mathtt{Iss}^P|$, and $\mathtt{Prf}^P$ is the set of dichotomous preference profiles on $\mathtt{Iss}^P$. There exists an agent $i \in \mathtt{Agn}^P$ such that i is almost decisive for some pair of issues.*

Proof. Let us proceed per absurdum assuming that there is no almost decisive agent. For condition **U**, there always exists for each pair of issues a set which is decisive for that pair, that is, $\mathtt{Agn}^P$. As a consequence, there always exists an almost decisive set for that pair. Let us take the smallest (possibly not unique) decisive set and call such set V. Given our hypothesis, such set cannot be a singleton. Let us divide V in three parts: a singleton $\{i\}$, a set $J = V - \{i\}$, and a set $K = \mathtt{Iss}^P - V$. Suppose the following holds: $y \prec_i x, z \approx_i y, z \prec_i x$; $y \prec_J x, y \prec_J z, z \approx_J x$; and $x \prec_K y, z \approx_K y, z \prec_K x$. As to the collective preference, since V is almost decisive one of the following must hold A) $y \prec x, y \prec z$; or B) $y \prec x, z \approx y$. However, if A) holds then J would be almost decisive contradicting the assumption that V was the smallest such set. It follows that B) is the only option. However, since B) is dichotomous it would also hold that $z \prec x$ which would make i almost decisive for x over z, against our assumption. We can argue symmetrically in the other cases. □

Proposition 6 follows directly from Lemmata 1 and 2.

As to Theorem 2 we just sketch the proof for lack of space. The same technique of the previous proof can be used by just redefining the notions of decisive and almost decisive sets of agents w.r.t. to single issues. A set $V \subseteq \mathtt{Agn}^P$ is *almost decisive* for formula ϕ (in symbols, $AD_V(\phi)$) iff: if $\forall i \in V, \phi \approx_i \bot$ (respectively, $\top$) and $\forall i \notin V, \phi \approx_i \top$ (respectively, $\bot$) then $\phi \approx \top$ (respectively, $\bot$). A set $V \subseteq \mathtt{Agn}^P$ is *decisive* for ϕ (in symbols, $D_V(\phi)$) iff: if $\forall i \in V, \phi \approx_i \top$ (respectively, $\bot$) then $\phi \approx_i \top$ (respectively, $\bot$). Obviously, for any $\phi \in \mathtt{Iss}^P$: $D_V(\phi)$ implies $AD_V(\phi)$. We need again two lemmata.

Lemma 3. *Let $\mathfrak{S}^P$ contain a set of issues $\mathtt{Iss}^P$ s.t. $\{p, q, p \wedge q\}$, and $\mathtt{Prf}^P$ is the set of Boolean preferences on $\mathtt{Iss}^P$. If there exists an individual $i \in \mathtt{Agn}^P$ such that $AD_V(\phi)$ for some pair ϕ then, under the conditions $\mathbf{U}^\top$ and $\mathbf{Sys}^\top$, i is decisive for any formula, that is, i is a dictator.*

Proof. We show that if i is almost decisive for any of the propositions in $\{p, q, p \wedge q\}$ then it is decisive for all of them. Let I denote $\mathtt{Agn}^P - \{i\}$. $[AD_i(p) \Rightarrow D(p \wedge q), D(q)]$ The proof proceeds like for Lemma 1 but makes use of systematicity. Assume $AD_i(p)$ and suppose the following holds: $p \approx_i \top, q \approx_i \top, p \wedge q \approx_i \top$ and $p \approx_I \bot$. In the collective preference, which must be Boolean, one of the following must hold: A) $p \approx \top, q \approx \top$; or B) $p \approx \top, q \approx \bot$. However, B cannot be the case because of $\mathbf{U}^\top$ and $\mathbf{Sys}^\top$. In fact, if $q \approx \bot$ then $q \approx_I \bot$ otherwise, for $\mathbf{U}^\top$, we should have $q \approx \top$. Consider a profile identical w.r.t. to q and $p \wedge q$ but s.t. $p \approx_i \bot$. Then by $\mathbf{U}^\top$ it follows that $p \approx \bot$, and by $\mathbf{Sys}^\top$ it follows that $q \approx \top$ since the two profiles are such that p is ranked $\top$ in the first iff q is ranked $\top$ in the second, which is impossible. Hence, A) is the only option, which proves the claim. We can then argue symmetrically for the other cases. □

The analogous of Lemma 2 can easily be proven, which completes the proof of the theorem.

Davide Grossi
ICR, University of Luxembourg
6, rue Coudenhove-Kalergi
1359 Luxembourg-Kirchberg, Luxembourg
Email: davide.grossi@uni.lu

Aggregating Referee Scores: an Algebraic Approach

Rolf Haenni

Abstract

This paper presents a quantitative solution to the problem of aggregating referee scores for manuscripts submitted to peer-reviewed conferences or scientific journals. The proposed approach is a particular application of the Dempster-Shafer theory to a restricted setting, from which an interesting algebraic framework results. The paper investigates the algebraic properties of this framework and shows how to apply it to the score aggregation and document ranking problems. Our scheme is intended to support the paper selection process of a conference or journal, not to replace it.

1 Introduction

This paper addresses a real-world problem of quantitative judgement aggregation, one that is omnipresent in the academic world and thus of great importance to all of us. We consider the typical situation of an editor or program committee, who is in charge of evaluating the manuscripts that have been submitted to a journal or conference for publication. In a competitive setting of a scientific conference, where the maximal number of accepted papers is limited, the problem then is to select the best papers from those submitted. For this, each submission is typically sent to 3 or 4 referees, who are asked to comment and score the paper in their report. This is the core of the so-called *peer review process*, which is a well-established academic procedure to guarantee high scientific standards and to prevent the dissemination of unwarranted claims or unacceptable interpretations.

Many journals and conferences ask the referees to quantitatively score the papers by a pair of values from respective scales, one that reflects the overall[1] quality of the paper and one that indicates the referee's own level of expertise or confidence. For the selection of the best papers, the editor or program committee faces then the problem of combining those scores to establish an overall ranking, from which the highest-ranked papers are accepted. The combination of such referee scores is a simple real-world judgement aggregation problem. Most of the existing convenient on-line conference management systems (e.g. ConfMaster, ConfTool, EasyChair, Linklings, OpenConf, Paperdyne, Start V2, WebChairing, etc.) are very rich in all kind of features for effortlessly accomplishing many complicated tasks of the peer review process, but they are usually very poor in providing automated decision support tools for the aggregation and ranking of referee scores. As a consequence, the score aggregation and paper ranking problems are still being solved manually today, and due to the many aspects and parameters to be taken into account, the resulting time-consuming procedure risks at producing bad quality results in form of unfair decisions. But what would be a reasonable procedure of combining the referee scores and establishing a ranking of peer-reviewed papers automatically?

1.1 Related Work

By describing a set of so-called *process patterns*, Nierstrasz gives an informal answer to the above question [26]. Examples of such patterns are "*Group papers according to their*

[1] Some journals and conferences ask to score various quality criteria independently of each other. The method presented in this paper is compatible with such multi-criteria scores, but we not treat them explicitly.

highest and lowest score" or "*Take care to identify papers with both extreme high and low scores*". For the scores, Nierstrasz proposes four quality categories A="*Good paper*" to D="*Serious problems*" and three levels of expertise X="*I am an expert*" to Z="*I am not an expert*".[2] Notice that Nierstrasz' pattern language has become something like the de facto standard for conferences in many computer science areas, and it is implemented in the conference management systems CyberChair [33] and Continue [23], and rudimentally in Confious [27], HotCRP [22], and MyReview [28]. The success of Nierstrasz' patterns is perfectly comprehensible from the pragmatic point of view of experienced program committee members, but from the more formal perspective of an expert in quantitative judgement aggregation or reasoning and decision making under uncertainty, they give the impression of being constituted on an ad hoc basis and may therefore seem a bit rudimentary.

A partial answer to the above question can be found in the early literature on probability from the late 17th and early 18th centuries [20, 30]. At that time, studying probability was often motivated by judicial applications, such as the reliability of witnesses in the courtroom, or more generally by the credibility of testimonies on past events or miracles. The first two combination rules for testimonies were published in an anonymous article [1]. One of them considers two independent witnesses with respective credibilities (frequencies of saying the truth) p_1 and p_2. If we suppose that they deliver the same report, they are either both telling the truth with probability $p_1 \cdot p_2$, or they are both lying with probability $(1-p_1)\cdot(1-p_2)$. Every other configuration is obviously impossible. The ratio of truth saying cases to the total number of cases,

$$\frac{p_1 \cdot p_2}{p_1 \cdot p_2 + (1-p_1)\cdot(1-p_2)}, \tag{1}$$

represents then the combined credibility of both witnesses. The more general formula for n independent witnesses of equal credibility p,

$$\frac{p^n}{p^n + (1-p)^n}, \tag{2}$$

has been mentioned by Laplace [24]. This formula is closely related to the *Condorcet Jury Theorem* discussed in social choice theory [2, 25]. Boole mentioned in [3] a similar formula that includes a prior probability of the hypothesis in question.

A recent article picks up these ancient ideas and turns them into a very general and flexible model of combining reports from partially reliable sources [15]. The generality of the model allows it to be applied to situations of incompetent or even dishonest witnesses, who may deliver highly contradictory testimonies. At its core, the model presupposes a non-additive measure of belief [13] in form of Dempster-Shafer belief functions [9, 29], but Laplace's and Boole's formulae are included as additive special cases. The model also includes various Bayesian approaches, which require a prior probability of the hypothesis in question to turn it into a corresponding posterior probability. The method discussed in this paper is another very particular case of the general model from [15].

1.2 Problem Formulation

The formal problem setting in this paper is the following. Let $\mathcal{D}$ be a set of submitted manuscripts (documents) and $\mathcal{R}$ a set of referees. We assume that the referees are anonymous and independent. The set of assigned referees for document $D \in \mathcal{D}$ is denoted by

[2] It should be emphasized here that Nierstrasz' categorization includes an explicit operational meaning, e.g. "*I will champion the paper*" for the category A="*Good paper*" and ditto for the other categories. The whole point of the pattern language is about how people will *behave* during a program committee meeting, which is why it is not directly applicable to a non-interactive setting such as journal paper reviewing. During PC meetings, however, the notion of championing can help to keep discussions focussed.

referees$(D) \subseteq \mathcal{R}$, that is *referees* : $\mathcal{D} \to \mathcal{P}(\mathcal{R})$ is a mapping from $\mathcal{D}$ to the power set of $\mathcal{R}$. Similarly, *documents*$(R) =$ *referees*$^{-1}(R) \subseteq \mathcal{D}$ denotes the set of documents assigned to referee $R \in \mathcal{R}$. If *referees*$(D) = \{R_1, \ldots, R_k\}$ is the set of referees assigned to a particular document D, then we assume to obtain a set of respective scores, *scores*$(D) = \{s_1, \ldots, s_k\}$, each of which being a pair $s_i = (q_i, e_i) \in [0,1] \times [0,1]$ of values between 0 and 1.[3] The value $q_i \in [0,1]$ is interpreted as the referee's estimate of the paper's overall quality and $e_i \in [0,1]$ as the referee's estimate of its own level of expertise, with the usual convention that higher values represent higher quality and expertise levels.

Given the above formal setting, this paper addresses the following interlinked problems of aggregating and ranking the given referee scores.

Aggregation. For each $D \in \mathcal{D}$, derive from *scores*(D) the document's combined overall score $s_D = (q_D, e_D) \in [0,1] \times [0,1]$.

Ranking. For a given set of combined overall scores, $\mathcal{S} = \{s_D : D \in \mathcal{D}\}$, determine a total preorder $\succeq$ according to which the documents in $\mathcal{D}$ are ranked (that is D_1 is preferred to D_2 iff $s_{D_1} \succeq s_{D_2}$).

The proposed process is thus a two-step procedure, in which the papers' overall scores appear as an intermediate result before the final ranking is established. Note that the ranking problem includes the decision problem of accepting/rejecting papers as a borderline case. For a single document, i.e. for $|\mathcal{D}| = 1$, we obtain another borderline case, one that corresponds to the situation of a single paper submitted to a journal.

As an alternative, it could also be assumed that each referee's set of assigned papers, *documents*$(R) \subseteq \mathcal{D}$, is first turned into a local ranking (or decision), from which the global ranking (or decision) is then established in a second step. Such a procedure is suggested in [11]. Its main advantage is the fact that scoring, with all its inherent problems such as the referees' diverging standards with judging the merits and weaknesses of each paper on a common scale, is no longer required. The whole problem of evaluating papers according to their overall quality is thus reduced to a ranking problem. Note that aggregating local rankings into a global ranking may become very difficult or even impossible if the average number of referees per paper is low. This is a consequence of the fact that rankings are usually less informative than corresponding sets of scores.

Another important advantage of the proposed two-step procedure is the possibility to repeat Step 1 with an updated set of scores. This may be necessary for papers with an unsatisfactory overall expertise level e_D. The ability to make such a distinction between papers reviewed by a group of experts from a paper reviewed by a group of non-experts is one of the main reason for the proposed 2-dimensional scores.

1.3 Overview

This paper proposes a solution to the two problems stated above. First we suggest a formal method for combining referee scores, and then we discuss a solution for projecting combined referee scores into a total order, from which the final document ranking results. The core of the suggested method is a particular application of what is known in the literature of uncertain reasoning as *Dempster's rule of combination*, a key concepts in the *Dempster-Shafer theory* (DST) of belief functions [9, 29]. Here we adopt Dempster's original interpretation of his theory as a generalization of classical probabilistic inference [10]. For this, we look at a single score $s_i = (q_i, e_i)$ of a peer-reviewed paper as a pair of respective probabilities.

[3] In practice, typical scales for referee scores are discrete sets such as $\{1, 2, \ldots, 10\}$ or {*very_poor*, *poor*, *medium*, *good*, *very_good*}. To make such cases compatible with our formal setting, we assume a mapping σ from the respective set into the unit interval $[0,1]$, e.g. $\sigma($*very_poor*$) = 0.1$, $\sigma($*poor*$) = 0.3$, etc.

Formally, this allows us to consider each score as a particular representation of the referee's *opinion* about the paper. Opinions are the key elements in Jøsang's theory of subjective logic [18, 19], a particular interpretation of DST. To deal with such opinions mathematically, we follow the algebraic setting proposed in [7, 17], in which the set of all possible opinions is considered as a commutative monoid with respect to Dempster's rule of combination.

In Section 2, we first give a short introduction to the Dempster-Shafer theory, and then we present the algebraic structure of the above-mentioned opinion calculus. This is the mathematical and computational foundation of the proposed score aggregation method presented in Section 3, where scores are interpreted as respective probabilities in a restricted Dempster-Shafer model. The conclusions in Section 4 close the paper.

2 The Opinion Calculus

In its original form [9, 10], the Dempster-Shafer theory proposes a particular application of probabilistic reasoning. Its main components are two sample spaces Ω and Θ, which are interlinked by a multi-valued mapping $\Gamma : \Omega \to \mathcal{P}(\Theta)$. A set $\Gamma(\omega) \subseteq \Theta$ is thus assigned to every element $\omega \in \Omega$. For a given probability space $(\Omega, \mathcal{F}, P)$, Dempster's theory shows how to carry the probability measure $P : \mathcal{F} \to [0, 1]$ defined over a σ-algebra $\mathcal{F}$ of subsets of Ω into a system of *lower* and *upper probabilities* over subsets $A \subseteq \Theta$,

$$\begin{aligned} P_*(A) &= P(\{\omega \in \Omega : \Gamma(\omega) \subseteq A\} \,|\, \{\omega \in \Omega : \Gamma(\omega) \neq \emptyset\}) \\ &= \frac{P(\{\omega \in \Omega : \emptyset \neq \Gamma(\omega) \subseteq A\})}{1 - P(\{\omega \in \Omega : \Gamma(\omega) = \emptyset\})} \end{aligned} \tag{3}$$

and

$$P^*(A) = 1 - P_*(A^c) = \frac{P(\{\omega \in \Omega : \Gamma(\omega) \cap A \neq \emptyset\})}{1 - P(\{\omega \in \Omega : \Gamma(\omega) = \emptyset\})}, \tag{4}$$

for which $P_*(A) \leq P^*(A)$ holds for all $A \subseteq \Theta$. A quadruple $(\Omega, P, \Gamma, \Theta)$ is sometimes called *hint* [21] or *Dempster space* [16].

Later, Shafer introduced a non-probabilistic interpretation of the Dempster's theory and suggested *belief* and *plausibility*, denoted by $Bel(A)$ and $Pl(A)$, as replacements for lower and upper probability [29]. Shafer's viewpoint and terminology has been adopted by many other authors, e.g. by Smets in his *Transferable Belief Model* [32]. They all depart from Dempster's original model by considering belief and plausibility functions over the so-called *frame of discernment* Θ that are not necessarily induced by an underlying probability space $(\Omega, \mathcal{F}, P)$ and its link over the multi-valued mapping Γ. There is an axiomatic system for such belief and plausibility functions [31], similar to Kolmogorov's system of axioms for probabilities. If Θ is finite, belief and plausibility functions are often expressed in terms of their underlying *mass function*,

$$m(A) = P(\{\omega \in \Omega : \Gamma(\omega) = A\}), \tag{5}$$

which is additive with respect to $\mathcal{P}(\Theta)$ and thus sums up to one over all $A \subseteq \Theta$. It is obvious that $m : \mathcal{P}(\Theta) \to [0, 1]$ is a true generalization of a classical probability mass function $p : \Theta \to [0, 1]$, and that Bel, Pl, and m are connected as follows:

$$Bel(A) = \frac{1}{1 - m(\emptyset)} \sum_{\emptyset \neq B \subseteq A} m(B), \quad Pl(A) = \frac{1}{1 - m(\emptyset)} \sum_{A \cap B \neq \emptyset} m(B). \tag{6}$$

Many variations of this general scheme have been proposed in the literature, but here we prefer to strictly follow Dempster's and Shafer's original views.

2.1 Opinions

To solve the particular problem of this paper, we restrict Dempster's original probabilistic model to a very particular case. First of all, we only consider finite sample spaces Ω, which allows us to replace the σ-algebra $\mathcal{F}$ by the power set $\mathcal{P}(\Omega)$ of Ω. Second, we only consider frames of discernment Θ of size two, e.g. $\Theta = \{H, \neg H\}$, where H and $\neg H$ denote complementary outcomes. This implies that $\{\emptyset, \{H\}, \{\neg H\}, \Theta\}$ is the codomain of the multi-valued mapping Γ. The essence of the whole structure $(\Omega, P, \Gamma, \Theta)$ can then be reduced to a simple pair $(b, d) \in [0, 1] \times [0, 1]$ with $b = Bel(\{H\})$, $d = Bel(\{\neg H\}) = 1 - Pl(\{H\})$, and therefore $b + d \leq 1$. In [7, 16, 17], such pairs are called *Dempster pairs* and the set of all such pairs is called *Dempster domain*. Corresponding additive triplets $\varphi_H = (b, d, i)$ with $i = 1 - b - d$ are sometimes called *opinions* about H [14, 18, 19].[4] As shown in Figure 1, opinons can be depicted as respective points in an equilateral triangle (2-simplex) called *opinion triangle* [18].[5] The three coordinates represent respective degrees of *belief* (probability assigned to "*H is true*"), degrees of *disbelief* (probability assigned to "*H is false*"), and degrees of *ignorance*[6] (probability assigned to "*I don't know*"). The correct mathematical term for the geometry of this picture is *barycentric coordinates* [4].

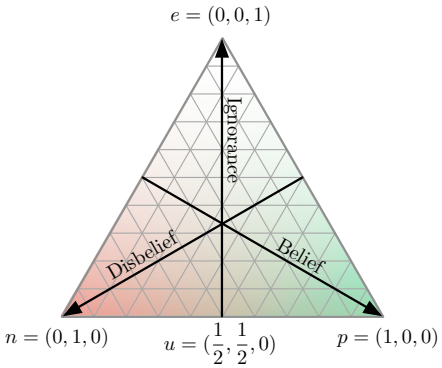


Figure 1: The opinion triangle with its three dimensions.

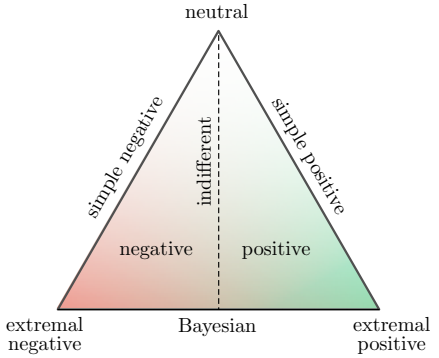


Figure 2: Various special types of opinions.

The three corners of the opinion triangle represent particular extreme cases. We adopt the terminology of [7, 17], i.e. $p = (1, 0, 0)$ and $n = (0, 1, 0)$ are called *extremal* and $e = (0, 0, 1)$ is called *neutral*. A general opinion (b, d, i) is called *positive* if $b > d$, and it is called *negative* if $b < d$. Positive opinions are located on the left hand side and negative opinions on the right hand side of the opinion triangle. (b, d, i) and (d, b, i) are regarded as *opposite* opinions, i.e. the opposite of a positive opinion is negative, and vice versa.

In the opinion triangle, the two regions of positive and negative opinions are separated by the central vertical line of *indifferent* opinions with $b = d$. Note that the neutral opinion e is indifferent. Other particular indifferent opinions are the points $u = (\frac{1}{2}, \frac{1}{2}, 0)$ at the bottom and $c = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ in the center of the triangle. Indifferent opinions are their own opposite.

Opinions are called *simple* (or *pure*) if either $b = 0$ or $d = 0$, i.e. $(b, 0, 1-b)$ is simple positive for $b > 0$ and $(0, d, 1-d)$ is simple negative for $d > 0$. Simple opinions are located on the left and the right edge of the triangle. Note that e, n, and p are simple. The opinions

[4] Later in [19], opinions are defined as quadruples (b, d, i, a) with an additional component a, the so-called *relative atomicity* (we do not need this in this paper).

[5] Sometimes, isosceles instead of equilateral triangles are used to visualize the Dempster domain [7, 16].

[6] In [19], Jøsang calls i degree of *uncertainty* rather than degree of ignorance, but the latter seems to be more appropriate and in better accordance with the literature.

$(b, 1-b, 0)$ at the bottom line of the triangle are called *Bayesian* (or *probabilistic*). Note that the extremal simple opinions p and n are also Bayesian, as well as the particular indifferent opinion u. All those particular types of opinions are shown in Figure 2.

2.2 Combining Opinions

One of the key components of the Demster-Shafer theory is a rule to combine two Dempster spaces $(\Omega_1, P_1, \Gamma_1, \Theta)$ and $(\Omega_2, P_2, \Gamma_2, \Theta)$ for a common frame of discernment Θ. Such a combination is usually denoted by symbols like $\otimes$ or $\oplus$ (here we prefer to use $\otimes$). If we assume P_1 and P_2 as being stochastically independent, then we naturally obtain

$$(\Omega_1, P_1, \Gamma_1, \Theta) \otimes (\Omega_2, P_2, \Gamma_2, \Theta) = (\Omega_1 \times \Omega_2, P_1 \cdot P_2, \Gamma_1 \cap \Gamma_2, \Theta) \tag{7}$$

for the combined structure [9, 21]. Translated into Shafer's terminology for two mass functions m_1 and m_2, we get what is known today as (unnormalized) *Dempster's rule of combination* or simply *Dempster's rule*:

$$m_1 \otimes m_2(A) = \sum_{B_1 \cap B_2 = A} m_1(B_1) \cdot m_2(B_2). \tag{8}$$

It is easy to see that Dempster's rule is commutative and associative, which means that the order in which opinions are combined is irrelevant. In the particular case of $\Theta = \{H, \neg H\}$ with two opinions $\varphi_1 = (b_1, d_1, i_1)$ and $\varphi_2 = (b_2, d_2, i_2)$, we know from [7, 12, 17] that Dempster's rule can be rewritten more compactly as

$$\varphi_1 \otimes \varphi_2 = \left(\frac{b_1 b_2 + b_1 i_2 + i_1 b_2}{1 - b_1 d_2 - d_1 b_2}, \frac{d_1 d_2 + d_1 i_2 + i_1 d_2}{1 - b_1 d_2 - d_1 b_2}, \frac{i_1 i_2}{1 - b_1 d_2 - d_1 b_2} \right), \tag{9}$$

and it includes (1) as special cases for $i_1, i_2 = 0$. This equation is the mathematical and computational basis for the proposed solution of the score aggregation problem in Section 3. Note that the combination of opposite extremal opinions, $p \otimes n$ or $n \otimes p$, is undefined.

2.3 The Opinion Monoid

The space of all possible opinions, $\Phi = \{(b, d, i) \in [0, 1]^3 : b + d + i = 1\}$, together with the particular form of Dempster's rule given in (9) forms an interesting algebraic structure. A thorough analysis of this structure is presented in [7, 17], where the set of all non-extremal opinions together with Dempster's rule is called *Dempster semigroup*. The extremal opinions are excluded to avoid the above-mentioned undefined combination. Here we pick up these ideas, but instead of excluding extremal opinions, we include $z = (1, 1, -1)$ as an additional opinion and call it *inconsistent*. The set of all such opinions, including the inconsistent one, is denoted by $\Phi_z = \Phi \cup \{z\}$, and $\otimes$ is extended by $p \otimes n = n \otimes p = z$. Note that z is absorbing with respect to $\otimes$, i.e. $z \otimes \varphi = \varphi \otimes z = z$ for all $\varphi \in \Phi_z$.

With this extension, the structure $(\Phi_z, \otimes)$ is closed under the operator $\otimes : \Phi_z \times \Phi_z \to \Phi_z$. From the commutativity and associativity of $\otimes$, it follows that $(\Phi_z, \otimes)$ is a *commutative semigroup*. Note that for $e = (0, 0, 1)$ we get $e \otimes \varphi = \varphi \otimes e = \varphi$ for all $\varphi \in \Phi_z$, i.e. $e = (0, 0, 1)$ is the (neutral) *identity element* of the combination. The structure $(\Phi_z, \otimes, e)$ is thus a *commutative monoid* with an absorbing *zero element* z. Since $\otimes$ is generally not invertible, $(\Phi_z, \otimes, e)$ is not a group. As is it a common practice in abstract algebra, we will refer to $(\Phi_z, \otimes, e)$ simply as Φ_z and call it the *opinion monoid*. Note that Φ_z has exactly three idempotent elements e, u, and z, i.e. $\varphi \otimes \varphi = \varphi$ only holds for $\varphi \in \{e, u, z\}$.

As pointed out in the classification of the previous subsection, Φ_z contains a number of interesting subsets. Some of them are again closed under combination andpreserve the

above-mentioned algebraic properties. Table 1 gives an overview of those sub-monoids with their respective identity and zero elements. Note that a non-extremal Bayesian opinion is invertible by combining it with its own opposite, i.e. $(b, 1-b, 0) \otimes (1-b, b, 0) = u$ holds for all $b \notin \{0, 1\}$. Mathematically speaking, the set of consistent non-extremal Bayesian opinions, $\Phi_0 \setminus \{p, n, z\}$, forms an commutative (abelian) group. Note further that the set $\Phi_0 \setminus \{z\}$ possesses a natural total order $\succeq_0$ defined by $(b_1, 1-b_1, 0) \succeq_0 (b_2, 1-b_2, 0)$ iff $b_1 \geq b_2$. This order is important in Section 3 to establish the document ranking.

Name	Notation	Definition	Identity	Zero
general (extended with z)	Φ_z	$\Phi \cup \{z\}$	e	z
non-negative	$\Phi_\geq$	$\{(b, d, i) \in \Phi : b \geq d\}$	e	p
simple non-negative	Φ_+	$\{(b, d, i) \in \Phi : d = 0\}$	e	p
non-positive	$\Phi_\leq$	$\{(b, d, i) \in \Phi : b \leq d\}$	e	n
simple non-positive	Φ_-	$\{(b, d, i) \in \Phi : b = 0\}$	e	n
indifferent	$\Phi_=$	$\{(b, d, i) \in \Phi : b = d\}$	e	u
Bayesian (extended with z)	Φ_0	$\{(b, d, i) \in \Phi : i = 0\} \cup \{z\}$	u	z

Table 1: Different algebraic sub-structures of the opinion monoid Φ_z with their respective identity and zero elements.

2.4 Transformations

As pointed out in [7, 17], the combination of a general opinion $\varphi \in \Phi_z$ with the uniform Bayesian opinion u defines a *homomorphism* $h : \Phi_z \to \Phi_0$, i.e. $h(\varphi_1 \otimes \varphi_2) = h(\varphi_1) \otimes h(\varphi_2)$ holds for all $\varphi_1, \varphi_2 \in \Phi_z$. We can thus use h to transform a general opinion $\varphi \in \Phi_z$ into a Bayesian opinion $h(\varphi) = \varphi \otimes u \in \Phi_0$. For $\varphi = (b, d, i) \in \Phi$ we can simplify (9) into

$$h(\varphi) = \left(\frac{1-d}{(1-b)+(1-d)}, \frac{1-b}{(1-b)+(1-d)}, 0 \right) = \left(\frac{1-d}{1+i}, \frac{1-b}{1+i}, 0 \right), \tag{10}$$

whereas $h(z) = z$ holds as usual. A more general version of this mapping is called *plausibility transformation* and $h(\varphi)$ is called *relative plausibility* [5, 6, 8]. Note that indifferent opinions always map into u, i.e. $h(\varphi) = u$ holds for all $\varphi \in \Phi_=$. This includes $h(e) = u$ as a special case. In the opinion triangle, applying h to a general opinion $\varphi \in \Phi$ means to intersect the straight line through φ and z with the bottom line of the opinion triangle [7, 17]. In other words, the intersection of the opinion triangle with the straight line through $\varphi_0 \in \Phi_0 \setminus \{z\}$ and z corresponds to the preimage $h^{-1}(\varphi_0) = \{\varphi \in \Phi : h(\varphi) = \varphi_0\}$ of φ_0. This geometric interpretation of h is illustrated in Figure 3.

A similar, but non-homomorphic transformation $g : \Phi_z \to \Phi_0$ results from applying a scheme similar to the one in (10). Instead of replacing the components b and d of a general opinion $\varphi = (b, d, i)$ by respective normalized plausibilities, the idea of

$$g(\varphi) = \left(\frac{b}{b+d}, \frac{d}{b+d}, 0 \right) \tag{11}$$

is to normalize b and d directly.[7] A general form of this mapping is called *belief transformation* and $g(\varphi)$ is called *relative belief* [6, 8]. Note that $g(e)$ is undefined in (11), but we can set $g(e) = u$ by default. As shown in Figure 4, applying g to $\varphi \in \Phi_z \setminus \{e\}$ means to intersect the straight line through the points φ and e with the bottom line of the opinion triangle. This geometric interpretation also holds for the special case $g(z) = u$.

[7] This transformation is a homomorphism with respect to the *disjunctive rule of combination* [7].

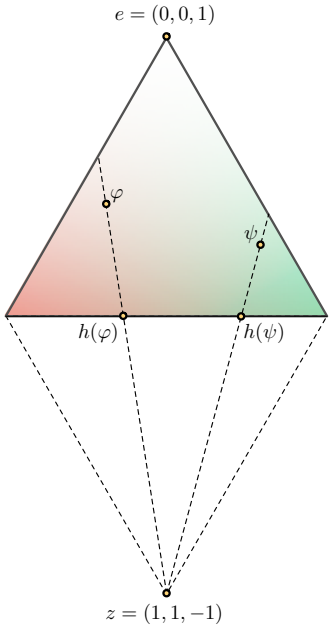


Figure 3: The homomorphic plausibility transformation h.

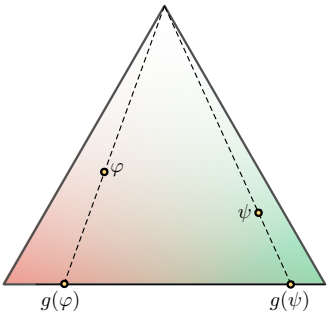


Figure 4: The belief transformation g.

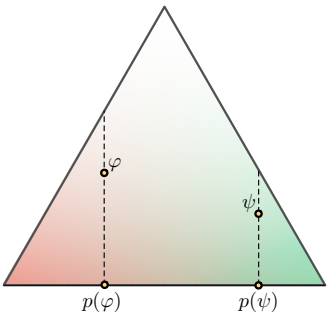


Figure 5: The pignistic transformation p.

Another interesting non-homomorphic transformation $p : \Phi_z \to \Phi_0$ redistributes the component i from a general opinion $\varphi = (b, d, i) \in \Phi_z$ equally among b and d,

$$p(\varphi) = \left(b + \frac{i}{2}, d + \frac{i}{2}, 0\right), \tag{12}$$

which includes $p(z) = u$ as a special case. In the opinion triangle, applying p to $\varphi \in \Phi$ means to project φ vertically onto the bottom line of Bayesian opinions. This is illustrated in Figure 5.[8] Note that p is a special case of what Smets calls *pignistic transformation* [32].

We prefer not to give any recommendation with regard to the question of which transformation to use, all of them have their respective advantages and disadvantages [5, 32]. In general, one can say that g tends to enforce existing degrees of belief and disbelief less cautiously than h, and that p lies somewhere in between.

3 Combining and Ranking Referee Scores

In this section, we use the algebraic investigation of the previous section as the mathematical and computational foundation for solving the problems of combining and ranking referee scores. First we show how to interpret the scores of a given document as respective opinions, one for each referee to which the document has been assigned. The independence assumption allows us then to apply the combination operator $\otimes$ defined in (9) to obtain the combined *group opinion* of all referees, from which the documents overall score is derived. Finally, we

[8] It is interesting to observe in Figure 3–5 that g, h, and p are special cases of a whole class of *symmetric transformations* obtained by intersecting the bottom line of the opinion triangle with a straight line through φ and $\varphi^i = (\frac{1-i}{2}, \frac{1-i}{2}, i)$ with $i \in \mathbb{R} \setminus [0, 1) \cup \{-\infty, +\infty\}$. In particular, we have $i = 1$ for g, $i = -1$ for h, and $i = \pm\infty$ for p.

explain how to use the proposed transformations g, h, or p to establish the final ranking, from which the highest-ranked documents are accepted.

3.1 Combining Referee Scores

As stated in Subsection 1.2, we first need to consider the problem of deriving from the set of *scores*(D) the document's combined overall score s_D. The general idea for this is to apply Dempster's rule to corresponding opinions. Recall that a score is a point $s = (q, e) \in [0,1]\times[0,1]$ in the unit square. The obvious question now is how to map such scores into opinions. Mathematically speaking, we are looking for a meaningful mapping $\Delta : [0,1]\times[0,1] \to \Phi$, which assigns a unique opinion $\Delta(s) \in \Phi$ to each score $s \in [0,1]\times[0,1]$. We can thus look at Δ as a transformation of the unit square into the opinion triangle.

To define such a transformation, we consider each referee as a partially reliable information source. Inspired by the general model of partially reliable information source in [15], we assume that reports of unreliable sources are entirely neglected. Intuitively, this is the case whenever a referee is not an expert for reviewing a particular paper. Note that $s = (q, e)$ delivers an estimate $e \in [0,1]$ of the referee's expertise level, i.e. if we assume the referee as being trustworthy with respect to giving such an estimate, then we may interpret e as the probability $P(\{E\}) = e$ of the referee being an expert for the document's topic. Similarly, we may interpret the quality estimate q as the conditional probability $P(\{Q\}|\{E\}) = q$ of the paper being a high-quality paper, given that the referee is an expert. With $\Omega_E = \{E, \neg E\}$ and $\Omega_Q = \{Q, \neg Q\}$ we denote respective sets of outcomes. This implies a probability space $(\Omega, \mathcal{P}(\Omega), P)$ with $\Omega = \Omega_E \times \Omega_Q = \{(E,Q), (E,\neg Q), (\neg E, Q), (\neg E, \neg Q)\}$ and

$$P(\{E,Q\}) = e \cdot q, \qquad P(\{\neg E, Q\}) = (1-e)\cdot q,$$
$$P(\{E,\neg Q\}) = e \cdot (1-q), \qquad P(\{\neg E, \neg Q\}) = (1-e)\cdot(1-q).$$

Consider now another space $\Theta = \{H, \neg H\}$ and let $\Gamma : \Omega \to \mathcal{P}(\Theta)$ be defined by $\Gamma(E,Q) = \{H\}$, $\Gamma(E,\neg Q) = \{\neg H\}$, and $\Gamma(\neg E, Q) = \Gamma(\neg E, \neg Q) = \Theta$. The idea is to adopt the referee's judgment with respect Q and $\neg Q$ whenever the referee is an expert, and to discard it otherwise. This defines a Dempster space $(\Omega, P, \Gamma, \Theta)$, from which we derive $Bel(\{H\}) = e \cdot q$ and $Bel(\{\neg H\}) = e \cdot (1-q)$. Finally, this leads to the requested transformation,

$$\Delta(s) = (e \cdot q, e \cdot (1-q), 1-e), \tag{13}$$

which maps points from the unit square into the opinion triangle, as illustrated in Figure 6. Note that all scores $(q, 0)$ are mapped into the neutral opinion $(0, 0, 1)$, whereas all scores $(q, 1)$ are mapped into Bayesian opinions $(q, 1-q, 0)$.

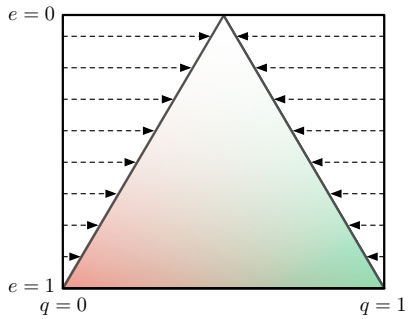


Figure 6: Transforming referee scores into opinions.

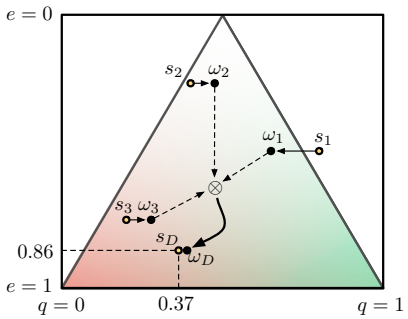


Figure 7: Combining three referee scores s_1, s_2, and s_3.

Let $scores(D) = \{s_1, \ldots, s_k\}$ be the set of scores for document D. Using the proposed transformation Δ, we can now compute the document's combined score s_D by

$$s_D = \Delta^{-1}(\Delta(s_1) \otimes \cdots \otimes \Delta(s_k)), \tag{14}$$

where Δ^{-1} denotes the inverse of Δ. Note that $\Delta^{-1}(\varphi) = (\frac{b}{1-i}, 1-i)$ is unique for all $\varphi = (b, d, i) \in \Phi$, except for the neutral opinion $(0, 0, 1)$. This is a mathematical imperfection, but it is of no importance for our particular application. Figure 7 shows three scores $s_1 = (0.8, 0.5)$, $s_2 = (0.4, 0.25)$, $s_3 = (0.2, 0.75)$, their combination $s_D = (0.37, 0.86)$, and corresponding opinions $\omega_i = \Delta(s_i)$ and $\omega_D = \omega_1 \otimes \omega_2 \otimes \omega_3 = \Delta(s_D)$.

3.2 Ranking Combined Referee Scores

Given the above solution for the score aggregation problem, we are now in possession of a set $\mathcal{S} = \{s_D : D \in \mathcal{D}\}$ of combined referee scores $s_D = (q_D, e_D)$, one for each document. The first thing to note here is the problem of documents with an unsatisfactory combined expertise level e_D. This means that the program committee or the editor was unable to assign the document to an appropriate referee. The usual procedure in such a case is to assign the paper to one or two additional referees. In our model, we may use a threshold $\gamma \in [0, 1]$ to define the set $\mathcal{D}_\gamma = \{D \in \mathcal{D} : e_D \leq \gamma\}$ of documents to be reassigned. The new referee scores are then combined with the existing ones to obtain an updated set of scores. This is a really important point when it comes to improve the quality of the review process, but time constraints usually do not allow more than one such iteration. Nierstrasz talks about the problem of low overall expertise levels in a pattern called *Identify Missing Champions* [26].

When all the updates are done and $\mathcal{S}$ is finally fixed, we reach the final stage of the review process, in which the decision about the accepted papers needs to be taken. An ideal basis for this decision would a be a ranking of the submitted documents, from which the highest-ranked submissions are accepted. We have seen in Section 3 that the set of Bayesian opinions is totally ordered, and that general opinions can be transformed into Bayesian opinions. The idea thus is to use the total order $\succeq_0$ of Bayesian opinions together with one of the proposed transformations to define an order with respect to Φ_z. More formally, we may consider three different orders $\succeq_g$, $\succeq_h$, and $\succeq_p$ for Φ, which are defined similarly by

$$\varphi \succeq_g \psi, \text{ iff } g(\varphi) \succeq_0 g(\psi), \quad \varphi \succeq_h \psi, \text{ iff } h(\varphi) \succeq_0 h(\psi), \text{ and } \varphi \succeq_p \psi, \text{ iff } p(\varphi) \succeq_0 p(\psi),$$

for all $\varphi, \psi \in \Phi_z$.[9] Note that $\succeq_g$, $\succeq_h$, and $\succeq_p$ are all *total preorders*, i.e. the antisymmetry property which is necessary for a total order does not hold. Nevertheless, we can use them to order the elements of $\mathcal{S}$, e.g. by

$$s_{D_1} \succeq_g s_{D_2}, \text{ iff } \Delta(s_{D_1}) \succeq_g \Delta(s_{D_2}), \text{ respectively } g(\Delta(s_{D_1})) \succeq_0 g(\Delta(s_{D_2})),$$

for g, and similarly for h and p. This is again a total preorder, which we can use to establish a document ranking for $\mathcal{D}$. Note that if $s_{D_1} \succeq_h s_{D_2}$ and $s_{D_2} \succeq_h s_{D_1}$ hold for two documents $D_1 \neq D_2$, which means that they have the same g-image in Φ_0, they are indistinguishable for $\succeq_g$ and thus receive the same rank. If necessary, such ties between equally ranked documents are broken at random.

[9] In the case of h, the order is only defined for Φ, i.e. it excludes the inconsistent opinion z obtained for two extreme opposite scores $(1, 1)$ and $(0, 1)$. To avoid this problem, either q should be restricted to $[0, 1)$, $(0, 1]$, or $(0, 1)$, or e should be restricted to $[0, 1)$. Another solution is to impose $h(z) = u$.

4 Conclusion

We have seen in this paper a solution for the score aggregation and document ranking problems. The key idea is to transform scores (q, e) into opinions (b, d, i), and to combine them by Dempster's rule. We have analyzed the underlying algebraic structures and properties. From various ways of projecting general opinions into the totally ordered set of Bayesian opinions, we can inherit the order and finally apply it to establish the final document ranking. The systematic formal analysis of this problem is the main contribution of this paper.

What is still missing today is the implementation of the proposed method in one of the existing conference management systems. A prototype implementation is available and can be tested at `http://www.iam.unibe.ch/~run/referee`, but it includes only the core of the proposed method with a very simple visualization. Nevertheless, it is interesting to observe that it almost perfectly reproduces some of Nierstrasz' classification patterns.

Another open issue is to apply and compare the recommended scheme empirically to real conference review data. It will certainly be interesting to observe whether real PC decisions are matched and to what extent. Note that we do not want to promote our method as a replacement for PC meetings or the discussions in editorial boards, but we think it could serve as a valuable decision support tool.

Acknowledgement

Thanks to Michael Wachter and Jacek Jonczy for careful proof-reading and to Milan Daniel and Oscar Nierstrasz for helpful discussions and comments. This research is partly supported by the SNF-project No. PP002–102652.

References

[1] Anonymous Author. A calculation of the credibility of human testimony. *Philosophical Transactions of the Royal Society*, 21:359–365, 1699.

[2] D. Black. *Theory of Committees and Elections*. Cambridge University Press, Cambridge, USA, 1958.

[3] G. Boole. *The Laws of Thought*. Walton and Maberley, London, 1854.

[4] O. Bottema. On the area of a triangle in barycentric coordinates. *Crux Mathematicorum*, 8:228–231, 1982.

[5] B. R. Cobb and P. P. Shenoy. On the plausibility transformation method for translating belief function models to probability models. *International Journal of Approximate Reasoning*, 41(3):314–330, 2006.

[6] F. Cuzzolin. Semantics of the relative belief of singletons. In *UncLog'08, International Workshop on Interval/Probabilistic Uncertainty and Non-Classical Logics*, number 46 in Advances in Soft Computing, pages 201–213, Ishikawa, Japan, 2008.

[7] M. Daniel. Algebraic structures related to the combination of belief functions. Technical Report 872, Academy of Sciences of the Czech Republic, Prague, Czech Republic, 2002.

[8] M. Daniel. On transformations of belief functions to probabilities. *International Journal of Intelligent Systems*, 21(3):261–282, 2006.

[9] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[10] A. P. Dempster. A generalization of Bayesian inference. *Journal of the Royal Statistical Society*, 30(2):205–247, 1968.

[11] J. R. Douceur. Paper rating vs. paper ranking. In *WOWCS'08, Workshop on Organizing Workshops, Conferences, and Symposia for Computer Systems*, 2008.

[12] M. Ginsberg. Non-monotonic reasoning using Dempster's rule. In *AAAI'84, 4th National Conference on Artificial Intelligence*, pages 112–119, Austin, USA, 1984.

[13] R. Haenni. Non-additive degrees of belief. In F. Huber and C. Schmidt-Petri, editors, *Degrees of Belief*. Springer, 2008.

[14] R. Haenni. Probabilistic argumentation. *Journal of Applied Logic*, 2008.

[15] R. Haenni and S. Hartmann. Modeling partially reliable information sources: a general approach based on Dempster-Shafer theory. *International Journal of Information Fusion*, 7(4):361–379, 2006.

[16] P. Hájek, T. Havránek, and R. Jiroušek. *Uncertain Information Processing in Expert Systems*. CRC Press, Boca Raton, USA, 1992.

[17] P. Hájek and J. J. Valdés. Generalized algebraic approach to uncertainty processing in rule-based expert systems (dempsteroids). *Computers and Artificial Intelligence*, 10:29–42, 1991.

[18] A. Jøsang. Artificial reasoning with subjective logic. In A. C. Nayak and M. Pagnucco, editors, *2nd Australian Workshop on Commonsense Reasoning*, Perth, Australia, 1997.

[19] A. Jøsang. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(3):279–311, 2001.

[20] J. Kohlas. Reliability of arguments. In E. von Collani, editor, *Defining the Science of Stochastics*, Sigma Series in Stochastics, pages 73–94. Heldermann, Lemgo, Germany, 2004.

[21] J. Kohlas and P. A. Monney. *A Mathematical Theory of Hints – An Approach to the Dempster-Shafer Theory of Evidence*, Springer, 1995.

[22] E. Kohler. Hot crap! In *WOWCS'08, Workshop on Organizing Workshops, Conferences, and Symposia for Computer Systems*, San Francisco, USA, 2008.

[23] S. Krishnamurthi. The CONTINUE server (or, how i administered PADL 2002 and 2003). In *PADL'03, 5th International Symposium on Practical Aspects of Declarative Languages*, LNCS 2562, pages 2–16, New Orleans, USA, 2003.

[24] P. S. Laplace. *Théorie Analytique des Probabilités*. Courcier, Paris, 3ème edition, 1820.

[25] Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'Imprimerie Royale, Paris, France, 1785.

[26] O. Nierstrasz. Identify the champion. In N. Harrison, B. Foote, and H. Rohnert, editors, *Pattern Languages of Program Design*, volume 4, pages 539–556. Addison-Wesley, 2000.

[27] M. Papagelis, D. Plexousakis, and P. Nikolaou. CONFIOUS: Managing the electronic submission and reviewing process of scientific conferences. In *WISE'05, 6rd International Conference on Web Information Systems Engineering*, LNCS 3806, pages 711–720, New York, USA, 2005.

[28] P. Rigaux. An iterative rating method: Application to web-based conference management. In *SAC'04, 19th Annual ACM Symposium on Applied Computing*, pages 1682–1687, 2004.

[29] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[30] G. Shafer. The early development of mathematical probability. In I. Grattan-Guinness, editor, *Companion Encyclopedia of the History and Philosophy of the Mathematical Sciences*, pages 1293–1302, London, U.K., 1993. Routledge.

[31] P. Smets. Quantifying beliefs by belief functions: An axiomatic justification. In R. Bajcsy, editor, *IJCAI'93: 13th International Joint Conference on Artificial Intelligence*, pages 598–603, Chambéry, France, 1993.

[32] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–234, 1994.

[33] R. van de Stadt. CyberChair: a web-based groupware application to facilitate the paper reviewing process. available on-line at http://www.borbala.com/cyberchair/wbgafprp.pdf, 2001.

Rolf Haenni

Bern University of Applied Sciences
CH-2501 Biel, Switzerland
Email: `rolf.haenni@bfh.ch`

and

University of Bern
CH-3280 Bern, Switzerland
Email: `haenni@iam.unibe.ch`

A Qualitative Vickrey Auction

Paul Harrenstein, Tamás Máhr, and Mathijs de Weerdt

Abstract

The negative conclusions of the Gibbard-Satterthwaite theorem—that only dictatorial social choice functions on three or more alternatives are non-manipulable—can be overcome by restricting the class of admissible preference profiles. A common approach is to assume that the preferences of the agents can be represented by *quasilinear utility functions*. This restriction allows for the positive results of the Vickrey auction and the Vickrey-Clarke-Groves mechanism. Quasilinear preferences, however, involve the controversial assumption that there is some commonly desired commodity or numeraire—money, shells, beads, etcetera—the utility of which is commensurable with the utility of the other alternatives in question. We propose a generalization of the Vickrey auction, which does not assume the agents' preferences being quasilinear but still has some of its desirable properties. In this auction a bid can be any alternative, rather than just a monetary offer. Such an auction is also applicable to situations where no numeraire is available, when there is a fixed budget, or when money is no issue. In the presence of quasilinear preferences, however, the traditional Vickrey auction turns out to be a special case. In order to sidestep the Gibbard-Satterthwaite theorem, we restrict the preferences of the agents. We show that this qualitative Vickrey auction always has a dominant strategy equilibrium, which moreover invariably yields a weakly Pareto efficient outcome, provided there are more than two agents.

The work in this paper is an improved presentation of the idea introduced by Máhr and de Weerdt (2007) on auctions with arbitrary deals.

1 Introduction

Although it may often seem otherwise, even nowadays money is not always the primary issue in a negotiation. Consider, for instance, a buyer with a fixed budget, such as a government issuing a request for proposals for a specific public project, a scientist selecting a new computer using a fixed budget earmarked for this purpose, or an employee organizing a grand day out for her colleagues. In such settings, the buyer has preferences over all possible offers that can be made to him. A similar situation, in which the roles of buyers and sellers are reversed, occurs when a freelancer offers his services at a fixed hourly fee. If he is lucky, several clients may wish to engage him to do different assignments, only one of which he can carry out. Needless to say, the freelancer might like some assignments better than others. In the sequel we consider the general setting which covers all of the examples above and in which we distinguish between an issuer of a commission—the government, the scientist, the employee, or the freelancer in the examples above—and a number of bidders.

In order to get the best deal, the issuer could ask for offers and engage in a bargaining process with each of the bidders separately. Another option would be to start a (reverse) auction. In this paper, we show that even without money, it is possible to obtain a reasonable outcome in this manner. We propose an auction protocol in which the dominant strategy for each bidder is to make the offer that, among the ones that are acceptable to her, is most liked by the issuer. We also show that if all bidders adhere to this dominant strategy a weakly Pareto optimal outcome results, provided there are three or more bidders.

To run such an auction without money the preferences of the issuer are made public. Observe that if a single good is sold in an auction with monetary bids it can be assumed to

be common knowledge that bidders prefer low prices to higher ones, and sellers higher to lower ones. Our protocol closely follows the protocol of a Vickrey, or closed-bid second-price, auction (Vickrey, 1961). First each bidder submits an offer. The winner is the bidder who has submitted the offer that ranks highest in the issuer's preference order. Subsequently the winner has the opportunity to select any other alternative as long as it is ranked at least as high as the second-highest offer in the issuer's preference order. This alternative is then the outcome of the auction.

In the next section some general notations and definitions from implementation theory are introduced, and in Section 3 we formally define the qualitative auction sketched above for the setting in which the bidders are indifferent between all outcomes where they do not win the auction. This makes that we can sidestep the negative conclusions of the impossibility result by Gibbard (1973) and Satterthwaite (1975). We prove that a dominant strategy equilibrium exists in the qualitative Vickrey auction, which moreover yields a weakly Pareto efficient outcome for all preference profiles with three or more bidders. The rest of that section concerns several other properties like weak monotonicity and incentive compatibility. We conclude the paper by relating our work to other general auction types such as multi-attribute auctions.

2 Definitions

In this section we review some of the usual terminology of mechanism design and fix some notations. For more extensive expositions the reader be referred to Moore (1992), Mas-Colell et al. (1995), and Shoham and Leyton-Brown (forthcoming).

Let N be a finite set of agents and Ω a set of alternatives or outcomes. The agents are commonly denoted by natural numbers. By a *preference relation $\succsim_i$ of agent i* we understand a transitive and total binary relation (that is, a weak order or a total preorder) on Ω, with $\succ_i$ and $\sim_i$ denoting its strict and indifferent part, respectively. We use infix notation and write $a \succsim_i b$ to indicate that agent i values alternative a at least as much as alternative b. It is not uncommon to restrict one's attention to particular subsets of preference relations on Ω, for instance, the sets of quasilinear preferences or single-peaked preferences on Ω. Be Θ_i such a class for each $i \in N$, we have Θ denote $\Theta_1 \times \cdots \times \Theta_n$. A *preference profile $\succsim$ in Θ (over Ω and N)* is a sequence $(\succsim_1, \ldots, \succsim_n)$ in $\Theta_1 \times \cdots \times \Theta_n$ associating each agent with a preference relation over Ω.

Given a preference profile Θ on Ω, an outcome ω in Ω is said to be *weakly Pareto efficient* whenever there is no outcome ω' in Ω such that all agents i strictly prefer ω' to ω. Outcome ω said to be *Pareto efficient* if there is no outcome ω' in Ω such that that ω' is weakly preferred to ω by all agents and strictly preferred by some.

A *social choice function (on Θ)* is a map $f\colon \Theta \to \Omega$ associating each preference profile with an outcome in Ω. A social choice function on Θ is said to be (weakly) Pareto efficient whenever $f(\succsim)$ is (weakly) Pareto efficient for all preference profiles $\succsim$ in Θ.

A *mechanism (or game form) M* on a set Ω of outcomes is a tuple $(N, S_1, \ldots, S_n, g)$, where N is a set of n agents, for each agent i in N, S_i is a set of strategies available to i, and $g\colon S_1 \times \cdots \times S_n \to \Omega$ is a function mapping each strategy profile s in $S_1 \times \cdots \times S_n$ on an outcome in Ω. A mechanism $(N, S_1, \ldots, S_n, g)$ is said to be *direct (on Θ)* if each agent's strategies are given by her possible preferences, that is, if $S_i = \Theta_i$ for each agent i in N. For Ω a set of outcomes, a pair $(M, \succsim)$ consisting of a mechanism M on Ω and a preference profile $\succsim$ on Ω we refer to as a *game (on Ω)*. With a slight abuse of terminology, we will also refer to functions $s_i\colon \Theta \to S_i$ as *strategies* and sequences $s = (s_1, \ldots, s_n)$ of such functions, one for each agent, as *strategy profiles*.

An *equilibrium concept* (or *solution concept*) associates each game with a subset of its

strategy profiles; the set of strategy profiles thus associated may depend on the preference profile. A mechanism M is said to *implement a social choice function f on Θ in an equilibrium concept C* whenever for all preference profiles $\succsim$ in Θ there is some $s^*(\succsim) \in C(M, \succsim)$ with $f(\succsim) = g(s^*(\succsim))$.

A direct mechanism $M = (N, \Theta_1, \ldots, \Theta_n, g)$ is said to be *truthful (or incentive compatible) in an equilibrium concept C* whenever for each preference profile $\succsim$ each agent i revealing her true preferences $\succsim_i$ is an equilibrium in $C(M, \succsim)$, that is, if $\succsim$ itself is in $C(M, \succsim)$.

If for a mechanism $M = (N, S_1, \ldots, S_n, g)$ and an equilibrium concept C, $C(M, \succsim)$ is nonempty for all preference profiles $\succsim$, we can associate with M a direct mechanism $M^* = (N, \Theta_1, \ldots, \Theta_n, g^*)$ where for each $\succsim$ in Θ we have $g^*(\succsim) = g(s^*(\succsim))$ for some selected equilibrium $s^*(\succsim)$ in $C(M, \succsim)$. Intuitively, M^* mimicks M by asking the agents to reveal their preferences, be it truthfully or untruthfully, calculating equilibrium strategies s_i^* in M for them given the revealed preferences $\succsim$ and returning the outcome $g(s_1^*(\succsim), \ldots, s_n^*(\succsim))$. Thus we find that a social choice function f being implementable in C implies it being truthfully implementable in C, a fact better known as the *revelation principle*.

In this paper we will be primarily concerned with *dominant strategy equilibrium*, which is extensively studied in the context of mechanism design (Dasgupta et al., 1979; Green and Laffont, 1979) and in terms of which also the infamous Gibbard-Satterthwaite theorem is formulated. For the purposes of this paper we say that s_i^* is a *dominant strategy for an agent i* in a game $(M, \succsim)$, whenever no matter which strategies the other agents adopt, i is not worse off playing s_i^* than any other of her strategies, that is, if for all strategy profiles $s \in S$ and all $t_i \in S_i$ we have

$$g(s_1, \ldots, s_{i-1}, s_i^*, s_{i+1}, \ldots, s_n) \succsim_i g(s_1, \ldots, s_{i-1}, t_i, s_{i+1}, \ldots, s_n).$$

A strategy profile $s^* = (s_1^*, \ldots, s_n^*)$ is then said to be a *dominant strategy equilibrium* if s_i^* is a dominant strategy for all agents i in N. The advantage of dominant strategy equilibrium is that it is very robust. The dominant strategies of an agent i do not depend on the preferences of the other agents, they can be calculated on the basis of i's preferences alone. Moreover, there seems to be no reason why agents would play a strategy that fails to be dominant if a dominant one is available. On the downside is the Gibbard-Satterthwaite theorem, which says that implementation in dominant strategy equilibrium allows only for social choice functions in which one of the players is a dictator if one does not impose restrictions on the agents' preference relations.

3 A Qualitative Vickrey Auction

In the setting we consider, a commission is issued and auctioned among a set N of n agents, henceforth called *bidders*. The commission can get a number of alternative implementations denoted by A, which for presentational purposes we assume to be finite.[1] The commission is then assigned to a bidder who commits herself to implement it in a particular way. Thus the outcomes of the auction are given by pairs (a, i) of alternatives $a \in A$ and bidders i in N, that is, $\Omega = A \times N$. Intuitively, (a, i) is the outcome in which i wins the auction and implements alternative a. For each bidder i in N we have Ω_i denote $A \times \{i\}$, the set of *offers* i can make. Obviously, each offer is also an outcome, rather, we have $\Omega = \bigcup_{i \in N} \Omega_i$. We assume each bidder to be indifferent between outcomes in which the commission is assigned to another bidder, that is, $\omega \sim_i \omega'$ for all bidders i in N and all outcomes ω and ω'

[1] The definitions and results of this paper can be extended so as to hold for infinite sets of outcomes as well, provided appropriate restrictions on the bidders' preferences are imposed. We believe, however, that doing so would technically complicate things while contributing only little to the conceptual content of this paper.

in $\Omega \setminus \Omega_i$. In what follows we have Θ_i denote the set of i's preference profiles over Ω which comply with this restriction.

If $(a,i) \succsim_i (x,j)$ for some alternative x and some bidder j distinct from i, outcome (a,i) is said to be *acceptable to* i, and *unacceptable to* i, otherwise. That is, an outcome ω is acceptable to bidder i if i values at least as much as any outcome in which she does not win the auction. Observe that if $i \neq j$, any outcome $(a,j) \in \Omega_j$ is acceptable to i. Finally, a preference profile $\succsim$ is said to be *positive* if each for each bidder i the set Ω_i contains at least one outcome (a,i) which i strictly prefers to losing the auction, that is, to any outcome not in Ω_i. Positive preference profiles could be argued for in contexts where a bidder is assumed not to partake in the auction if she is at best indifferent between winning and losing.[2]

Let $\geq$ be a linear (that is, a transitive, total and anti-symmetric) order over the *outcomes* Ω. The *qualitative Vickrey auction on* $\geq$ is defined then by the following protocol. First, the order $\geq$ is publicly announced. For $\omega \geq \omega'$ we say that *outcome* ω *is ranked at least as high as outcome* ω' *in* $\geq$. Then, each bidder i submits a secret *offer* $(a,i) \in \Omega_i$ to the auctioneer. The bidder i^* who submitted the offer ranked highest in $\geq$ is declared the winner of the auction. Observe that ties are precluded because of the linearity of $\geq$. Finally, i^* may choose from among her own offers in Ω_{i^*} any outcome that is ranked at least as high as the offer that ranks *second highest* in $\geq$ among all the ones submitted. The outcome she chooses is then the outcome of the auction. The winner's initial offer is witness to the fact that such an outcome always exists.

Example 1 *Let* $N = \{1,2,3\}$ *and* $A = \{a,b,c,d\}$. *Let us further suppose that the order* $\geq$ *on the alternatives is lexicographic, that is,*

$$(a,1) > (a,2) > (a,3) > (b,1) > \cdots > (c,3) > (d,1) > (d,2) > (d,3).$$

Suppose the three bidders 1, 2, *and* 3 *submit the offers* $(c,1)$, $(a,2)$ *and* $(d,3)$, *respectively. Bidder* 2 *then emerges as the winner, as* $(a,2) > (c,1) > (d,3)$. *Since* $(c,1)$ *is the second-highest offer, bidder* 2 *may now choose from the outcomes* $(a,2)$ *and* $(b,2)$, *these being the only outcomes in* Ω_2 *that rank higher than* $(c,1)$. *In case bidder* 2 *prefers* $(b,2)$ *to* $(a,2)$ *she would only do well selecting* $(b,2)$, *which would then also be the outcome of the auction.*

For different orders $\geq$ on the outcomes, the qualitative Vickrey auction can obviously yield different outcomes. So, actually, we have defined a class of auctions. With a slight abuse of terminology we will nevertheless speak of *the* qualitative Vickrey auction if the respective order $\geq$ can be taken as fixed. At first we will consider $\geq$ an extraneous feature of the auction. Later we will come to consider the case in which $\geq$ represents the preferences of the issuer of the commission.

The traditional second-price or Vickrey auction, in which a single item is allocated, is a special case of the above protocol, when the alternatives are taken to be monetary bids for a single good, the bidders have quasilinear preferences over the outcomes and $\geq$ represents the natural order over monetary bids—ranking higher bids higher than lower ones—together with a deterministic tie-breaking rule.[3] Since from each offer the bidder's entire preference relation can be derived, the traditional Vickrey auction could be considered a direct mechanism. Moreover, being a special case of the VCG mechanism, it is incentive compatible in dominant strategies.

[2] In a similar vein, one could introduce a *zero outcome* 0, which represents the possibility of no transaction taking place. A bidder i could also offer 0, which would intuitively mean that i refrains from participating in the auction. Such a zero outcome, however, brings along a number of intricacies, which lie beyond the scope of this paper.

[3] Not all tie-breaking rules $\tau : 2^{\Omega} \to \Omega$, however, can be represented by $\geq$. *E.g.*, if τ is such that $\tau(\omega_1,\omega_2) = \omega_1$, $\tau(\omega_2,\omega_3) = \omega_2$ and $\tau(\omega_1,\omega_3) = \omega_3$, it cannot be represented by an order $\geq$. Moreover, we also assume the number of possible offers in the Vickrey auction to be arbitrarily large but finite.

The qualitative Vickrey auction, however, is not a direct mechanism, as from an offer the full preference relation of a bidder cannot be derived in general. As such incentive compatibility is not a concept that directly applies to it. Instead we prove the existence of a dominant strategy equilibrium $s^*(\succsim)$ for each preference profile $\succsim$ in Θ. Thus, the qualitative Vickrey auction implements a social choice function f^*, which is defined such that for all preference profiles $\succsim$ in Θ, $f^*(\succsim)$ is the outcome of the equilibrium $s^*(\succsim)$. We will then study the formal properties of this social choice function.

Intuitively, the classic Vickrey auction is truthful because an bidder's monetary offer only determines whether she turns out to be the winner, but not what price she has to pay if she does. Things are much similar in the qualitative Vickrey auction. Again, the bidder's offer determines whether she emerges as the winner, but the range of alternatives from among which she may choose is decided by the second-highest offer.

A strategy for a bidder i in the qualitative Vickrey auction consists of an offer (a, i) in Ω_i along with a contingency plan which outcome to choose from among the outcomes in Ω_i that are ranked higher than the second-highest offer submitted in case i happens to win the auction. Any such strategy may depend on a preference profile $\succsim$ in Θ. We call a strategy for i *adequate* if it satisfies the following properties:

(i) the offer i submits is the outcome in Ω_i that is ranked highest in $\geq$, and that is still acceptable to i,

(ii) in case Ω_i contains no outcomes acceptable to her, i submits the outcome in Ω_i that is ranked lowest in $\geq$,

(iii) in case i wins the auction, she selects one of the outcomes in Ω_i she values most among those that are ranked higher than the second-highest offer submitted.

Given a preference profile $\succsim$ items (i) and (ii) completely determine the offer i is to submit, but (iii) leaves some room for flexibility when i's preferences over Ω_i contain indifferences. Also observe that whether an offer is acceptable to a bidder i can be read off immediately from i's preference relation and does not depend on the preferences of the other bidders or other extraneous features.

Example 1 (continued) *Let the preferences of the three bidders* 1, 2 *and* 3 *be given by the following table, where higher placed outcomes are more preferred.*

1	2	3
$(c, 1)$	$(d, 2)$	$(x, i) \notin \Omega_3$
$(d, 1)$	$(b, 2)$	$(a, 3)$
$(x, i) \notin \Omega_1$	$(a, 2)$	$(d, 3)$
$(b, 1)$	$(x, i) \notin \Omega_2$	$(c, 3)$
$(a, 1)$	$(c, 2)$	$(b, 3)$

If the bidders 1, 2 *and* 3 *were all to play an adequate strategy, they would offer* $(c, 1)$, $(a, 2)$ *and* $(d, 3)$*, respectively, since these are for 1 and 2 their highest-ranking acceptable offer and for 3 the lowest-ranking offer overall. In this case* $(b, 2)$ *would be the outcome of the auction, because bidder 2 is the winner and may select any alternative ranked above* $(c, 1)$*. It might be worth observing that it can happen that, if all of her offers are unacceptable to her, a bidder adhering to the strategy offers her* least *preferred outcome. Bidder* 3*, for instance, would do so if the outcomes* $(b, 3)$ *and* $(d, 3)$ *had been interchanged in her preference order.*

We are now in a position to prove that the bidders' adequate strategies are dominant in the qualitative Vickrey auction.

Proposition 1 *In the qualitative Vickrey auction and given a preference profile $\succsim$ in Θ, all adequate strategies for a bidder i are dominant.*

Proof: Let i be an arbitrary bidder and $s(\succsim)$ an arbitrary adequate strategy for i. First assume that there are no outcomes in Ω_i that are acceptable to i and that i adheres to $s_i(\succsim)$ submitting the lowest ranked offer in Ω_i, denoted by (a_0^i, i). If i loses the auction, some other bidder i^* ends up winning the auction and chooses some offer (a^*, i^*) in Ω_{i^*} as the eventual outcome. Observe that (a^*, i^*) is acceptable to i and among her most preferred outcomes. If i wins the auction, she may choose among *all* outcomes in Ω_i and, following $s_i(\succsim)$ she will select one that she likes best. Any other offer she could make would still make her win the auction and leaving her the same range of outcomes to choose from. So, obviously, in both cases, $s_i(\succsim)$ is a dominant strategy.

For the remainder of the proof we may assume that there are outcomes in Ω_i which are acceptable to i. Let (a^i, i) denote the highest-ranked offer in Ω_i that is still acceptable to i, that is, the offer i would make if she follows the adequate strategy $s_i(\succsim)$. First assume that submitting (a^i, i) would make i lose the auction, that is, that some other bidder i^* would win the auction by offering (a, i^*) and choose (a^*, i^*) as the eventual outcome. Now consider any other offer (a', i^*) in Ω_i which i could submit. Obviously, if (a', i^*) were also a losing offer, i^* would still win the auction and i would be indifferent between the outcome i^* would then choose and (a^*, i^*). On the other hand, if (a', i) would make i win the auction, we have $(a', i) \geq (a, i^*)$, rendering (a, i^*) the second-highest offer. Then, i has to choose from among the outcomes in Ω_i ranked higher than (a, i^*). All of these outcomes, however, are unacceptable to i, that is, $(a^*, i^*) \succ_i \omega$ for all $\omega \in \Omega_i$ with $\omega \geq (a, i^*)$. Thus, also in this case we may conclude that $s_i(\succsim)$ is a dominant strategy for i.

Finally, assume that i wins the auction by offering (a^i, i) and that (b, j) is the second-highest offer. Let (a^*, i) be the outcome she chooses as her most preferred outcome among the outcomes in Ω_i that are ranked higher than (b, j). Then, $(a^i, i) \geq (a^*, i) > (b, j)$, because any outcome in Ω_i ranked higher than (a^i, i) is unacceptable to i. Obviously, $(a^*, i) \succsim_i \omega$ for any outcome $\omega \notin \Omega_i$. For any other winning offer, the second-highest offer would remain the same and so does the set of outcomes from which i may choose. Thus, i would do no better than by offering (a^i, i) as prescribed by $s_i(\succsim)$. On the other hand, if i were to submit a losing offer, some outcome $\omega \notin \Omega_i$ would result. Since $(a^*, i) \succsim_i \omega$, again i would have done better by offering (a^i, i). Hence, $s_i(\succsim)$ is a dominant strategy for i. □

Among the adequate strategies of a bidder i one stands out, namely, the one in which she selects from her most preferred outcomes that ranked higher than the second highest, the one that is ranked highest. For each preference profile $\succsim$ in Θ we denote this strategy by $s_i^*(\succsim)$. Let further s^* be the strategy profile such that $s^*(\succsim) = (s_1^*(\succsim), \ldots, s^*(\succsim))$ for each preference profile $\succsim$. Then, in virtue of Proposition 1, $s^*(\succsim)$ is a dominant strategy equilibrium for each $\succsim$ in Θ. Accordingly, the qualitative Vickrey auction on $\geq$ implements the social choice function $f_{\geq}^*$, which is such that for all preference profiles $\succsim$ in Θ, $f_{\geq}^*(\succsim)$ equals the outcome the strategy profile $s^*(\succsim)$ gives rise to. If $\geq$ is clear from the context we omit the subscript $\geq$ in $f_{\geq}^*$.

We are now in a position to define a *direct* mechanism $M^* = (N, \Theta_1, \ldots, \Theta_n, g^*)$ such that N are the bidders participating in the qualitative Vickrey auction we are considering, Θ_i the possible preference relations over Ω (restricted as in the beginning of this section), and g^* such that for all $\succsim$ in $\Theta_1 \times \cdots \times \Theta_n$ we have $g^*(\succsim) = f^*(\succsim)$.

Proposition 2 *The direct mechanism M^* truthfully implements the social choice function f^*.*

Proof: That M^* truthfully implements f^* is an almost immediate consequence of Proposition 1 by an argument much similar to that for the revelation principle. □

It is quite possible that, given a preference profile $\succsim$, if all bidders play an adequate (and hence dominant) strategy, the outcome (a^*, i^*) of the qualitative Vickrey auction is unacceptable to i^* although some submitted offers (a, i) were acceptable to the respective bidder i. To appreciate this consider once more Example 1 but now suppose that the bidders' preferences are such that all offers are unacceptable to them, apart from $(d, 2)$, which is acceptable to bidder 2. Then, bidder 1 would win the auction and be forced to select some outcome $(x, 1)$ that is unacceptable to her. This could, and probably should, be considered a serious weakness. Fortunately, this defect can easily be remedied in the direct mechanism M^* by selecting the winner from the bidders i with acceptable outcomes among their set Ω_i of possible offers, if such bidders exist. The problem can obviously also be sidestepped by assuming all preferences to be *positive*, that is, if for each bidder i the set Ω_i contains at least one acceptable outcome which i strictly prefers to losing the auction.

3.1 Pareto efficiency

The generalized Vickrey auction fails to be *(strongly) Pareto efficient among the bidders*, in the sense that for some preference profiles there could be an outcome (a^{**}, j) that is weakly preferred by all bidders over the dominant equilibrium outcome (a^*, i^*), and strictly preferred by some.

Proposition 3 *For any order $\geq$ on the outcomes, there is a preference profile for which the outcome of the qualitative Vickrey auction on $\geq$ is not Pareto efficient among the bidders.*

Proof: Let $\geq$ be any order on the outcomes and let (a, i) be the *lowest* ranked outcome therein. Now define the preference profile $\succsim$ such that for all bidders j distinct from i all outcomes in Ω_j are unacceptable to j and that (a, i) is the only outcome in Ω_i that i strictly prefers to losing the auction. Obviously, there is no way in which (a, i) can be the outcome of the auction. Still, (a, i) Pareto dominates any other outcome (a^*, i^*) with $i^* \neq i$: bidder i^* strictly prefers (a, i) to (a^*, i^*) whereas all other bidders are at least indifferent. □

In contrast to strong Pareto efficiency, *weak Pareto efficiency among the bidders* is satisfied almost trivially. A mechanism is weakly Pareto efficient if there are no preference profiles and orders $\geq$ such that some outcome is *strictly* preferred over the dominant equilibrium outcome by all bidders. If there are three or more bidders, for any two outcomes (a, i) and (b, j) there is some bidder k distinct from both i and j and thus $(a, i) \sim_k (b, j)$. In words, bidder k will never strictly prefer any outcome where she is not a winner. In the case with only two (distinct) bidders, say i and j, we have $(a, i) \sim_j (b, i)$ and $(a, j) \sim_i (b, j)$ for all $a, b \in A$. The only way, moreover, in which it can happen that both $(a, i) \succ_i (b, j)$ and $(a, i) \succ_j (b, j)$ is that (a, i) is acceptable to i and (b, j) unacceptable to j. However, (b, j) can turn out the dominant strategy equilibrium outcome only if j has no acceptable offers at all, which is a rather uninteresting borderline case.

Thus far, we have assumed the order $\geq$ to have been given externally. The order $\geq$ could of course also be construed as the preference relation of an additional bidder with a interest in the outcome of the auction, for instance, the issuer of the commission. Extending the concepts of Pareto efficiency so as to include the preferences of this new party, we find that the qualitative Vickrey auction is both weakly and strongly Pareto efficient provided that the preferences of each bidder i are positive and linear over Ω_i. Linearity can be dropped if we consider the direct mechanism M^*.

Proposition 4 *The qualitative Vickrey auction is strongly Pareto efficient among the bidders and $\geq$, if the preferences of each bidder i are positive and linear over Ω_i.*

Proof: Let (a^*, i^*) be a dominant strategy equilibrium outcome of the qualitative Vickrey auction. Having assumed the preferences to be positive, (a^*, i^*) is acceptable to i^*. We now show that (a^*, i^*) is not dominated by any other outcome. Consider an arbitrary outcome (a, i) in Ω distinct from (a^*, i^*). Without loss of generality we may assume that $(a, i) > (a^*, i^*)$. If $i = i^*$, then $(a^*, i^*) \succ_{i^*} (a, i)$ by linearity and the observation that otherwise, i^* would have not have selected (a^*, i^*). On the other hand, if $i \neq i^*$, the outcome (a, i) is ranked higher in $\geq$ than the second-highest offer. As such (a, i) is not acceptable to i, whereas (a^*, i^*) is. Hence, $(a^*, i^*) \succ_i (a, i)$. In either case, (a, i) does not Pareto dominate (a^*, i^*) strongly. □

3.2 Monotonicity

Another property of the social choice function implemented by the qualitative Vickrey auction is that of mononicity. A social choice function f on Ω is said to be *(weakly) monotonic on* Θ if $f(\succsim) = f(\succsim')$ for any preference profiles $\succsim$ and $\succsim'$ in Ω that only differ in that the social choice $f(\succsim)$ under $\succsim$ is possibly moved up in the individual preference orders $\succsim_i'$. In other words, is for all bidders i in N and all outcomes ω and ω' distinct from $f(\succsim)$, $\omega \succsim_i \omega'$ if and only if $\omega \succsim_i' \omega'$ and $f(\succsim) \succsim_i \omega$ implies $f(\succsim) \succsim_i' \omega$, then $f(\succsim) = f(\succsim')$. Intuitively, weak monotonicity captures the desirable property that if the social choice ω^* becomes more preferred by some or more bidders while the bidders' preferences over the other outcomes stay the same, ω^* remains the social choice. A mechanism is said to be weakly monotonic if the social choice functions it implements are weakly monotonic.

For the qualitative Vickrey auction we have imposed the restriction on the individual preferences that a bidder is indifferent between any outcome in which she does not win. In case there are two or more alternatives or more than two bidders, this makes that a loser i of the auction cannot move the outcome (a^*, i^*) up in his preference order, keeping all her other preferences intact, without violating this restriction. Hence, for weak monotonicity on Θ we only have to consider preference profiles that only differ in that the outcome (a^*, i^*) moves up in the preferences of the winner. We then find that the qualitative Vickrey auction is indeed weakly monotonic.

Proposition 5 *The qualitative Vickrey auction is weakly monotonic.*

Proof: If there is only one alternative and no more than two players the proof is trivial. For any other case consider two preference profiles $\succsim$ and $\succsim'$ in Θ and let (a^*, i^*) be the outcome of the auction if the bidders' preferences are given by $\succsim$. Without loss of generality we may assume that $\succsim_i$ and $\succsim_i'$ are identical for all bidders i distinct from i^*. Also assume that $\succsim_{i^*}$ and $\succsim_{i^*}'$ only differ in that (a^*, i^*) is moved up in $\succsim_{i^*}'$. We now show that (a^*, i^*) is also the outcome of the auction if the bidders' preferences are given by $\succsim'$. Observe that for all bidders distinct from i^* the sets of acceptable outcomes given $\succsim_i$ and $\succsim_i'$ remain the same. Hence, the highest-ranked offer (a, i) submitted by any bidder distinct from i^* will be identical given either $\succsim$ or $\succsim'$. Now either (a^*, i^*) is acceptable in $\succsim$ if and only if (a^*, i^*) is in $\succsim'$, or (a^*, i^*) is unacceptable in $\succsim$ but acceptable in $\succsim'$. In the former case, the offer by i^* given $\succsim'$ will be identical to her offer given $\succsim$. In the latter case i^* will offer (a^*, i^*) when the preferences are given by $\succsim'$. In either case i^* also wins the auction for $\succsim'$. Morever, (a^*, i^*) is one of the outcomes among those ranked higher in $\geq$ than (a, i) that i^* prefers most. By moving (a^*, i^*) up in i^*'s preference order, this remains the case and (a^*, i^*) will also be the outcome of the auction if the preferences are given by $\succsim'$. □

A social choice function f is said to be *strongly monotonic on* Θ if $f(\succsim) = f(\succsim')$ for all preference profiles $\succsim$ and $\succsim'$ in Θ such that $f(\succsim) \succsim_i \omega$ implies $f(\succsim) \succsim_i' \omega$ for all bidders i and all outcomes ω. This is a very strong property that is satisfied by hardly any reasonable

social choice function. It is therefore not very surprising that the qualitative Vickrey auction fails to be strongly monotonic as well, as witness the following example involving two bidders and three outcomes.

Example 2 *Let $\geq$ be given by $(a,1) > (a,2) > (b,1) > (b,2) > (c,1) > (c,2)$ and the preference profiles $(\succsim_1, \succsim_2)$ and $(\succsim'_1, \succsim_2)$ as follows.*

1	$1'$	2
$(c,1)$	$(c,1)$	$(b,2)$
$(b,1)$	$(x,i) \notin \Omega_1$	$(a,2)$
$(x,i) \notin \Omega_1$	$(b,1)$	$(c,2)$
$(a,1)$	$(a,1)$	$(x,i) \notin \Omega_2$

Bidder 1 and bidder 2 then offer $(b,1)$ and $(a,2)$, respectively, so that bidder 2 wins the auction and the outcome is $(a,2)$. However, moving $(a,2)$ up in bidder 1's preference order, together with $(b,2)$ and $(c,2)$ so as to comply with the restriction set on preference profiles, and leaving bidder 2's preferences intact results in the profile $(\succsim'_1, \succsim_2)$. Now, however, bidder 1 submits the losing offer $(c,1)$, leaving bidder 2 in a position to choose her most preferred outcome $(b,2)$.

3.3 Incentive compatibility for the issuer

So far we have assumed that the preference order of the issuer is publicly known, like the fact that a seller likes to get a higher price. In some settings however, this order $\geq$ may not be common knowledge. Therefore, we should also investigate whether the proposed mechanism is incentive compatible for the issuer as well. Unfortunately, we can show that this is not the case, leaving an open problem for future work to investigate how much the issuer can profit by lying.

Consider the following case where the mechanism is not incentive compatible for the issuer. As always, the winner can select an alternative that is equally or more preferred than the second-highest offer in the publicly known ordering. Suppose that there is an alternative in this set she strictly prefers to her own offer. By definition, this alternative is less preferred by the issuer than the highest offer. Had the issuer manipulated its order by moving the second highest offer up and position it right under the winner's offer, the winner would not have had any other choice than to accept her original offer.

For example, take the preferences and the offers from Example 1. Suppose the issuer moves the alternative $(c,1)$ up in its order to the spot between $(a,2)$ and $(a,3)$. In that case the dominant strategies for the bidders would still lead to the same offers, and the winner would still be bidder 2 with her offer $(a,2)$, but now she is only allowed to choose among the offers higher than or equal to $(c,1)$, which leaves $(a,2)$ as the only acceptable alternative. This outcome is better for the issuer than $(b,2)$, which was the outcome based on his true preference order.

4 Extensions and variants

In this section we consider a number of extensions and variants of the ideas underlying the qualitative Vickrey auction.

4.1 Other auction types

To start with, similar results on incentive compatibility and Pareto-efficiency can be obtained for the English auction in a straightforward manner. In this setting the auctioneer accepts

only bids in increasing order of the global ordering until no bidder is interested anymore. The dominant strategy for a bidder i is then to offer the highest acceptable alternative in her preference order that is higher in $\geq$ than the current accepted bid. The effect of this strategy is equivalent to the dominant strategy described earlier for the qualitative Vickrey auction: the winner is the bidder that has an acceptable offer that is highest in $\geq$ and the winning alternative is not dominated by any acceptable offer by any other bidder.

The qualitative auction protocol can also be rephrased for Dutch auctions, or first-price sealed bid auctions, but those are not incentive compatible. But then, neither are traditional variants of these auctions, when preferences are assumed to be quasilinear.

4.2 Multi-attribute auctions

The qualitative Vickrey auction does not assume that preferences of bidders can be expressed as quasilinear utility functions. This can be a feature for applications where preferences cannot easily be expressed in terms of money. Similar considerations play an important role in the related field of multi-attribute auctions. In a multi-attribute auction the good is defined by a set of attributes which can take different values. A bid consist of a value for each attribute and a price. Che (1993) analyzed situations where a bid consists of a price and a quality attribute, and proposed first-price and second-price sealed-bid auction mechanisms. His work was extended by David et al. (2002) for situations where the good is described by two attributes and a price. They analyzed the first-price sealed-bid, and English auction, and derived strategies for bids in a Bayesian-Nash equilibrium. In addition, they studied a setting where the issuer can also strategize, and they showed when and how much the issuer can profit from lying about his valuations of the different attributes. The main difference from our work is that in their approach the preferences of the auctioneer (issuer) and the bidders are related: better for the bidder means generally worse for the auctioneer.

Parkes and Kalagnanam (2005) concentrated on iterative multi-attribute reverse English auctions. Here prices of attribute-value combinations (a full specification of the good) are initially set high, and bidders submit bids on some attribute-value combinations to lower the prices. The auction finishes when there are no more bids. Such auctions allow the bidders to have any (non-linear) cost structure, and the authors claim that myopic best-response bidding—that is, the strategy always to bid a little bit below the current ask price—results in an ex-post Nash equilibrium for bidders, and that the auction then yields an efficient outcome. One of the main differences with our approach, besides theirs proposing an iterative protocol and using an ex-post Nash equilibrium as solution concept, is that they use quasilinear utility functions. To the best of our knowledge, such a restriction on the preferences of the issuer and the bidders being weakly inverse is essential to all of the existing work on (multi-attribute) auction mechanisms.

5 Discussion

In this paper we showed that there is another way of dealing with the impossibility theorem by Gibbard (1973) and Satterthwaite (1975) besides requiring quasilinear utility functions. For settings where there is only one winner, all that is required is that all bidders are indifferent between all outcomes where they are not the winner. We proposed a protocol for settings where the preference order of the issuer is publicly known, in a way similar to the public knowledge that sellers prefer high prices and buyers low prices. This protocol is called the qualitative Vickrey auction since it can be seen as a generalization of the Vickrey auction to a setting without quasilinear utility functions.

We defined a class of dominant strategies for this qualitative auction and saw that it is weakly Pareto efficient in the resulting equilibrium, provided there are more than three

bidders. We also found that the social choice function implemented by the qualitative Vickrey auction is weakly, but not strongly, monotonic. Furthermore, we showed that the mechanism is not incentive compatible for the issuer. We also briefly discussed the relation of the qualitative Vickrey auction to other auction types. Still, there are a number of interesting questions left unanswered regarding the properties of qualitative mechanisms such as the one presented here.

We would like to show how much worse off the bidders can be if the issuer turns out to be malicious. Another direction stems from the observation that we defined qualitative Vickrey auctions as a class of mechanisms, some of which are dictatorships, for instance when all outcomes with a particular winner are ranked above all outcomes where another bidder wins. It would be interesting to see precisely under which conditions on the issuer's order $\geq$ the qualitative Vickrey auction is not dictatorial. Also, we are interested in the properties of the qualitative Vickrey auction if additional restrictions on $\geq$ and the class of preference orders are imposed, for instance, if the preference orders of the bidders are assumed to be the weak inverse of $\geq$. Finally, we are interested in other qualitative generalizations of quasilinear mechanisms, for example of online auctions (Hajiaghayi et al., 2005).

Acknowledgements

We are indebted to Felix Brandt, Kevin Leyton-Brown and three anonymous referees for inspiring discussions and valuable comments. This work is partially supported by the Technology Foundation STW, applied science division of NWO, and the Ministry of Economic Affairs of the Netherlands as well as by the Deutsche Forschungsgemeinschaft under grants BR 2312/3-1 and BR 2312/3-2.

References

Y.K. Che. Design competition through multidimensional auctions. *RAND Journal of Economics*, 24:668–680, 1993.

P. Dasgupta, P. Hammond, and E. Maskin. The implementation of social choice rules: Some general results on incentive compatibility. *The Review of Economic Studies*, 46:185–216, 1979.

E. David, R. Azoulay-Schwartz, and S. Kraus. Protocols and strategies for automated multi-attribute auctions. In *Proceedings of the First Joint Conference on Autonomous Agents and Multiagent Systems*, pages 77–85, 2002.

A. Gibbard. Manipulation of voting schemes: a general result. *Econometrica*, 41(4):587–602, July 1973.

J.R. Green and J.-J. Laffont. *Incentives in Public Decision Making*. North-Holland, Amsterdam, 1979.

M. Hajiaghayi, R. Kleinberg, M. Mahdian, and D. Parkes. Online auctions with re-usable goods. In *In Proc. 6th ACM Conf. on Electronic Commerce (EC-05)*, pages 165–174, 2005.

T. Máhr and M.M. de Weerdt. Auctions with arbitrary deals. In V. Marek, V. Vyatkin, and A.W. Colombo, editors, *HoloMAS 2007*, volume 4659 of *LNAI*, pages 37–46. Springer-Verlag Berlin Heidelberg, 2007. ISBN 978-3-540-74478-8.

A. Mas-Colell, M. D. Whinston, and J. R. Green. *Microeconomic Theory*. Oxford University Press, Inc., 1995.

J. Moore. Implementation, contracts, renegotiations in environments with complete information. In J.J. Laffont, editor, *Advances in Economic Theory*, chapter 5, pages 182–282. Cambridge University Press, 1992.

D.C. Parkes and J. Kalagnanam. Models for iterative multiattribute procurement auctions. *Management Science*, 51(3):435–451, 2005. Special Issue on Electronic Markets.

M. A. Satterthwaite. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10:187–217, 1975.

Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, Cambridge, forthcoming.

W. Vickrey. Counter speculation, auctions, and competitive sealed tenders. *Journal of Finance*, 16(1):8–37, 1961.

Paul Harrenstein
Theoretische Informatik
Institut für Informatik
Ludwig-Maximilians-Universität München
80538 Munich, Germany
Email: `paul.harrenstein@ifi.lmu.de`

Tamás Máhr
Algorithmics Group
Faculty of Electrical Engineering, Mathematics, and Computer Science
Delft University of Technology
2628 CD, Delft, The Netherlands
Email: `t.mahr@tudelft.nl`

Mathijs de Weerdt
Algorithmics Group
Faculty of Electrical Engineering, Mathematics, and Computer Science
Delft University of Technology
2628 CD, Delft, The Netherlands
Email: `M.M.deWeerdt@tudelft.nl`

How to Rig Elections and Competitions

Noam Hazon and Paul E. Dunne and Sarit Kraus and Michael Wooldridge

Abstract

We investigate the extent to which it is possible to rig the agenda of an election or competition so as to favor a particular candidate in the presence of imperfect information about the preferences of the electorate. We assume that what is known about an electorate is the *probability* that any given candidate will beat another. As well as presenting some analytical results relating to the complexity of finding and verifying agenda, we develop heuristics for agenda rigging, and investigate the performance of these heuristics for both randomly generated data and real-world data from tennis and basketball competitions.

1 Introduction

The impossibility theorems of Arrow and Gibbard-Satterthwaite are perhaps the most famous results of social choice theory [1]. At first sight, these results seem to tell us that the design of social choice mechanisms is a quixotic enterprise, since any mechanism we care to define will fail to satisfy a basic "reasonableness" axiom, or will be prone to strategic manipulation. However, in a seminal 1989 paper, Bartholdi, Tovey, and Trick argued that the fact that a voting rule in susceptible to manipulation *in theory* does not mean that is manipulable *in practice*, since it may be computationally infeasible (NP-complete or worse) to determine *how* to manipulate an election optimally [3]. Since the work of Bartholdi *et al*, many researchers have investigated the extent to which voting mechanisms can be manipulated (see, e.g., [6] for a recent collection of papers on this and related topics). Most work on the manipulation of voting procedures has considered the manipulation of elections by *voters*; specifically, the strategic misrepresentation of preferences in order to bring about a more favored outcome. However, manipulation is also possible by election officers – those responsible for organising an election, in which context it is sometimes called "control" [4].

In this paper, we consider election control by rigging the ballot agenda in order to favor a particular candidate. It is well-known that some sequential pairwise majority elections may be rigged in this way – see, e.g., [5, p.177] and [9]. In such an election, we fix an ordering of the candidates (the voting agenda), so that the first two candidates in the ordering will be in a simple majority ballot against each other, with the winner then going on to face a ballot against the third candidate, and so on, until the winner of the final ballot is the overall winner. If we know the preferences of the electorate – or more specifically, who would win in every possible ballot – then it may be possible to fix the election agenda to the benefit of a favored candidate [8].

However, the assumption that we know exactly how a voter would vote in any given ballot is very strong, and ultimately unrealistic. It ignores the possibility of strategic voting, for one thing, but more generally, the preferences of voters will *not* be public – we will have at best only *partial* information about them. In light of this, the present paper considers the extent to which it is possible to rig an election agenda (and, more generally, running orders for competitions) in the manner described above in the presence of *imperfect* information. We assume that an election officer knows the *probability* that a given candidate will beat another in a pairwise ballot. This probability may be obtained from opinion polls, in the case of governmental elections or similar; or it may be from form tables, in the case of sporting competitions. Given this, we investigate the problem of rigging an election agenda in the more realistic setting of imperfect information. We study the complexity of

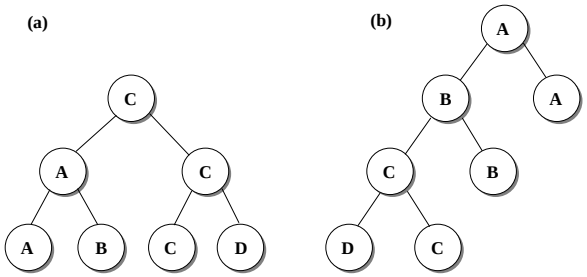


Figure 1: Organising a series of pairwise majority ballots into trees (a) and linear orders (b).

finding and verifying agendas, and present heuristics for agenda rigging. We investigate the performance of these heuristics for both randomly generated data sets and real-world data sets from tennis and basketball competitions. Finally, it is worth clarifying the motivation for this work. We are, of course, *not* advocating election manipulation, or trying to develop techniques to make it easier! If we can identify cases where election manipulation is easy in practice (even if it is hard in theory), then we can use this information to design elections so as to avoid the possibility of manipulation.

2 Preliminary Definitions

Although our problem is closely related to voting procedures, the way we choose to model the imperfect information here places it on a wider scope which includes also sports tournament. We assume we have a set $\Omega = \{\omega_1, \ldots, \omega_n\}$ of *outcomes* or *candidates* – the things the voters are trying to decide over. We are concerned with settings in which we simply want to select one outcome from Ω; more specifically, we are concerned with procedures to select such an outcome through a *sequence of pairwise simple majority ballots*. Such selection procedures are used in both political elections and sporting competitions. We often summarise voter preferences in a *majority graph*, $G \subseteq \Omega \times \Omega$, where $(\omega, \omega') \in G$ means that ω would beat ω' in a pairwise simple majority ballot. An election officer will not usually know the preferences of individual voters, but they may know the *probability* that one candidate will beat another. If all probabilities are 0 or 1 then we say the scenario is one of *perfect information*, otherwise it is one of *imperfect information*. An *imperfect information ballot matrix* M is an $|\Omega| \times |\Omega|$ matrix of probabilities, such that if $M[\omega_i, \omega_j] = p$, then in a ballot between ω_i and ω_j, candidate ω_i will win with probability p. We require that $M[\omega_i, \omega_j] + M[\omega_j, \omega_i] = 1$.

The most obvious way to organise a series of pairwise majority elections is in a *voting tree* In Figure 1(a), we see how ballots between candidates A, B, C and D may be organised into such a tree. The idea is that candidates A and B face each other in a ballot, while candidates C and D face each other in a ballot. The winner of the first ballot (A in this case) then faces the winner of the second (C), and the winner of this third ballot (C) is declared the overall winner. The voting tree in Figure 1 is said to be *fair* because it is *balanced*, and as a consequence every possible overall winner would have to win the same number of ballots. The task of the election officer may thus be seen as generating such a binary tree, with candidates Ω allocated to leaves of the tree. More formally, a ballot tree for an n candidate election is an n leaf tree, $T(V, E)$, having a distinguished root $r(T)$ and in which every node with the exception of the leaves has exactly two children. An *instantiated agenda* for a binary ballot tree is a bijective mapping $\alpha : \Omega \leftrightarrow leaves(T)$, which associates a candidate with each leaf node in T. Given $\langle T, \alpha, M \rangle$ the *outcome evaluation*

of T with respect to α and M is a mapping $\eta : V \rightarrow [0,1]^n$ such that for any $v \in V$, $\eta(v) = \langle v_1, \ldots, v_n \rangle$ if and only $Pr[\ \omega_i$ is the winner at $v\ ] = v_i$. Thus $\eta(r(T))$ will contain at index i the probability that ω_i will be the overall winner of T.

The opportunity for manipulation arises in such a setting because the election officer can, for example, place a favored candidate against candidates it is likely to beat. If we relax the fairness constraint, then the possibilities for an election officer to manipulate the election increase. Figure 1(b) shows a rather unfair voting tree; in fact, it defines a *linear order* (C, D, B, A) for the candidates, with the first ballot taking place between C and D, the winner facing B, and so on, until the winner of the final ballot (A in this case) is the overall winner. The unfairness arises because it is possible for a candidate to win overall despite only participating in one ballot (as is the case in Figure 1). We will denote linear voting orders (i.e., permutations of Ω) by $\pi, \pi', \ldots$.

A different model of imperfect information was studied in [8], where it was assumed that for each voter we have a correct but incomplete model of their preference relation. For this incomplete information setting, they considered questions such as whether there was some completion of the incompletely known preferences and some voting tree for the candidates that would make a desired candidate a winner. Roughly, our aim in this paper is to study election manipulation in much the same way as [8], but with the probability model of imperfect information described above, rather than the incomplete profile model.

3 Voting with a Linear Order of Ballots

We start by considering *linear* voting orders. Given such an order $\pi = (\omega_1, \ldots, \omega_n)$ and outcome $\omega^* \in \Omega$, the probability of ω^* being the overall winner of π is denoted by $Pr[w(\pi) = \omega^* \mid M]$, and is given as follows. For a voting order $\pi = (\omega_{i_1}, \omega_{i_2}, \ldots, \omega_{i_n})$ with $\omega^* = \omega_{i_k}$, (i.e., the preferred winner occurs k'th in the order π),

$$P[w(\pi) = \omega^* \mid M] = \varphi \times \psi$$

where

$$\varphi = \left(\prod_{j=k+1}^{n} M[\omega_{i_k}, \omega_{i_j}] \right)$$

$$\psi = \left(\sum_{j=1}^{k-1} P[w(\omega_{i_1}\omega_{i_2}\ldots\omega_{i_{k-1}}) = \omega_{i_j} \mid M] \times M[\omega_{i_k}, \omega_{i_j}] \right)$$

That is, in order for ω^* to emerge as the winning candidate, ω^* must defeat every candidate put forward *later* in the voting order, *and* succeed against the eventual winner of the voting order formed by the *earlier* candidates. We will show in the next section that the probability of a candidate being a winner in any given voting tree can be computed in time $O(n^3)$; for linear orders, we can improve this to $O(n^2)$; we omit the proof.

What else can we say about voting with linear orders? First, we can make precise, and prove correct, the intuition that there is no benefit to going early; a candidate can only benefit by going late in a voting order. While this seems intuitive, the proof of the following lemma, which we omit, is surprisingly involved.

Lemma 1 *Given $\langle M, \Omega \rangle$ let ω_i, ω_j be distinct members of Ω. For every voting order $\pi_1\pi_2$ of $\Omega \setminus \{\omega_i, \omega_j\}$, it holds that $Pr[w(\pi_1\omega_j\omega_i\pi_2) = \omega_i] \geq Pr[w(\pi_1\omega_i\omega_j\pi_2) = \omega_i]$.*

The immediate corollary is as follows.

Corollary 1 *For any candidate $\omega^* \in \Omega$, if there is a voting order, π such that $Pr[w(\pi) = \omega^*] \geq p]$, then there is a voting order π' such that $Pr[w(\pi') = \omega^*] \geq p]$ and in which ω^* is the final candidate to run.*

Proof: Given π with $Pr[w(\pi) = \omega] \geq p$, apply Lemma 1 repeatedly to move ω later in the voting order until it is the final candidate. ▫

Note that the equivalent result *does not* hold for trees. It is possible to construct an example in which a candidate wins with probability 1 in a fair tree, but loses with probability 1 if we convert the tree to a linear order with the same candidates order and place the candidate last. Even though it faces fewer ballots the candidate has a strictly smaller probability of winning.

It is also natural to ask what the probability is that a particular candidate is a *possible winner*, that is, if there is *some* permutation of Ω which would make ω^* the winner with non-zero probability. The decision problem POSS-WIN takes as its instance a triple $\langle M, \Omega, \omega^* \rangle$, which is accepted iff ω^* is a possible winner of $\langle M, \Omega \rangle$.

Theorem 1 POSS-WIN *is polynomial time decidable.*

Proof: (Outline) Convert all the non-zero probabilities to one, and apply the techniques of [8] to check whether ω^* is a possible winner. ▫

The general problem of determining whether there exists any agenda which gives a named candidate at least a certain probability of winning is hard to classify, and so we will analyse a restricted version of the problem, as follows. When we think informally about rigging an agenda, we tend not to think just in terms of the agenda, but also in terms of the *specific outcomes* that we want the agenda to lead to. So, we might think in terms of "if I put A up against B, then B wins and goes up against C, and C wins..." and so on. Here, we have not just the agenda (ABC) but also the outcomes of the ballots (B wins the first; C the second; ...). Here, we call these structures – which include the agenda for the ballots together with the intended outcomes – a *run*. A run has the form $r : \omega_1, \omega_2 \xrightarrow{\omega_2} \omega_3 \xrightarrow{\omega_3} \cdots \xrightarrow{\omega_{k-1}} \omega^* \xrightarrow{\omega^*}$ where ω_1 and ω_2 are the candidates up against each other in the first ballot, ω_2 is the intended winner of this ballot, and so on, until the final ballot is between ω_{k-1} and ω^*, in which we intend the winner – and hence overall winner – to be ω^*. Computing the probability that this run will result in our desired candidate ω^* winning is simple – it is the value: $Pr[w(\omega_1, \omega_2) = \omega_2] \times Pr[w(\omega_2, \omega_3) = \omega_3] \times \cdots \times Pr[w(\omega^{k-1}, \omega^*) = \omega^*]$. We denote this value for a run r by $Pr[r \mid M]$. So, in the *weak imperfect information rigged agenda* (WIIRA) problem, we are given a set of candidates Ω, an imperfect information ballot matrix M, a favored candidate $\omega^* \in \Omega$, and a probability p. We are asked whether there exists a run r, in which the overall winner is ω^*, such that $Pr[r \mid M] \geq p$.

Theorem 2 WIIRA *is* NP*-complete.*

Proof: (Summary) A standard "guess and check" algorithm gives membership in NP. For hardness, we reduce the k-HCA problem on tournaments [2, p.46]: we are given a tournament $G = (V, E)$ (i.e., a complete digraph such that $(\omega, \omega') \in E$ iff $(\omega', \omega) \notin E$) and a subset $E' \subseteq E$, and we are asked whether G contains a Hamiltonian cycle containing all edges E'. We create an instance of WIIRA as follows. The outcomes will be the vertices of G together with a new vertex, $v_\perp$. Given a tournament $G = (V, E)$ and required edge set E', we create a probability matrix so that G contains a cycle with E' iff we can create an ordering of vertices satisfying the property given above. For each $(u, v) \in E'$ we set $M[u, v] = 1$. For each $(u, v) \in E \setminus E'$ we set $M[u, v] = 1 - \frac{1}{10^{|V|}}$. We then set the target probability to be $(1 - \frac{1}{10^{|V|}})^{|V|-|E'|}$. We are after a cycle, so we need to select a vertex (call it $v_\top$) to act as

the source of the cycle and our new vertex, $v_\perp$, will act as the sink in the cycle; if we have an arc $(u, v_\top) \in E'$ then we define $M[u, v_\perp] = 1$, while if $(u, v_\top) \in E \setminus E'$ then we define $M[u, v_\perp] = 1 - \frac{1}{10^{|V|}}$; if for any vertex $u \in V$ we have not yet defined a value for $M[u, v_\perp]$ then define $M[u, v_\perp] = 0.5$. We then ask whether we can rig the agenda for $v^\top$ to win with probability greater than the target. ▫

4 Voting with a Tree of Ballots

We now focus on a more realistic setting, in which ballots are organised into a binary tree. There are several obvious questions to ask here. For example, the most obvious decision problem with respect to rigging an election as follows: Given a set of outcomes Ω, an imperfect information ballot matrix M, a favored candidate $\omega^* \in \Omega$ and a probability p, does there exist a voting tree T with labeling $\alpha : \Omega \leftrightarrow \mathit{leaves}(T)$ such that ω^* wins in this setting with probability at least p? It is not obvious even that this problem is in NP, since to compute the probability that a given candidate wins overall, we must consider every possible ordering of wins arising from a given tree structure: in any given ballot, there are two outcomes, in contrast to the perfect information case. However, the following result implies the problem is in NP.

Theorem 3 *Let $T(V, E)$ be any binary ballot tree, α a labeling of the leaves of T by candidates Ω, and M a ballot matrix for Ω. The outcome evaluation of T with respect to α and M is computable in $O(n^3)$ arithmetic operations.*

Proof: Consider the following algorithm.

1. $\eta(x) = \mathit{unlabelled}$ for each $x \in V(T)$
2. For each leaf, x, of T, $\eta(x) = \langle x_1, \ldots, x_n \rangle$ with $x_i = 1$ if $\alpha(x) = \omega_i$ and $x_i = 0$ otherwise.
3. **repeat**
   a. Let z be any node of T with children x and y such that $\eta(x) \neq \mathit{unlabelled}$ and $\eta(y) \neq \mathit{unlabelled}$. Let $\langle x_1, \ldots, x_n \rangle$ denote $\eta(x)$ and, similarly, $\langle y_1, \ldots, y_n \rangle$ denote $\eta(y)$. Compute $\eta(z) = \langle z_1, \ldots, z_n \rangle$ using

$$z_i \quad := \quad \begin{cases} x_i \ \sum_{j=1}^{n} \ y_j \ M[\omega_i, \omega_j] & \text{if} \quad x_i > 0 \\ y_i \ \sum_{j=1}^{n} \ x_j \ M[\omega_i, \omega_j] & \text{if} \quad y_i > 0 \\ 0 & \text{if} \quad x_i = y_i = 0 \end{cases}$$

4. **until** every v has $\eta(v) \neq \mathit{unlabelled}$

This algorithm can be improved in some cases; e.g., if we know that the graph is *fair*, then we can use an optimised version to take account of this, reducing the overall time complexity to $O(n^2)$. ▫

Whether this result is matched by NP-hardness remains open, but we can obtain a related hardness result, as follows. Given a set of candidates Ω and an imperfect information ballot matrix M, a *fair tree run* r_T is a balanced tree, T, a labeling α of the leaves of T by candidates Ω, and a labeling of every other node of T by the candidate that has the higher probability to win the ballot between the node's children (If a node's children have equal probability, we select one arbitrarily). The motivation for this is just as in the linear order case; we tend to think on the agenda in terms of the *specific outcomes* that we want the

agenda to lead to. The probability that a fair tree run will result in our favored candidate ω^* winning is simply the multiplication of the winning probabilities of the candidates in all the ballots induced by T. We denote this value by $Pr[r_T \mid M]$. So, in the *weak imperfect information rigged agenda for balanced trees* (WIIRA-BT) problem, we are given a set of candidates Ω, an imperfect information ballot matrix M, a favored candidate $\omega^* \in \Omega$, and a probability p. We are asked whether there exists a fair tree run r_T, in which the overall winner is ω^*, such that $Pr[r_T \mid M] \geq p$.

Theorem 4 WIIRA-BT *is* NP*-complete.*

Proof: (Summary) Membership of NP is immediate. As proved by [8], given a complete and weighted majority graph (also called a weighted tournament) G, every balanced voting tree with candidate ω^* at its root has a corresponding binomial tree with root ω^* which covers all the nodes in G – this binomial tree describes all the ballots between the candidates. We will prove our hardness result with respect to the following problem, a variation of the problem from [8, Th. 4], which is easily shown to be NP-complete: *Given a weighted tournament $G = (V, E)$, a candidate ω^* and a cost bound k, check whether there exist a balanced voting tree in which ω^* wins, such that the sum of the edges of the corresponding binomial tree is at least k.* Let denote $\min_E(\max_E)$ as the minimum (maximum) weight of the edges in E. Given an instance of the previous problem we create an instance of WIIRA-BT with a graph $G' = (V, E')$ such that for every edge $(u, v) \in E$ with a weight $w(u, v)$, we create the same edge in E' with the weight

$$2^{\frac{w(uv)-|\max_E|}{|\min_E-|\max_E||}}.$$

This conversion ensures that $0.5 \leq w(u, v) \leq 1$ in E'. For every edge $(u, v) \in E'$, we create an opposite edge (v, u) with the weight $1 - w(u, v)$. The candidate for our problem is also ω^*, and the winning probability p is

$$2^{\frac{k-(n-1)\times|\max_E|}{|\min_E-|\max_E||}}.$$

The overall number of ballots in a balanced voting tree is $n - 1$ so $k \leq (n - 1) \times |\max_E|$ and $0 \leq p \leq 1$ as required. If we denote the edges of a binomial tree for ω^* in G with a cost bound k as $x_1, x_2, \ldots, x_{n-1}$, then:

$$x_1 + x_2 + \cdots + x_{n-1} \geq k$$

$$(x_1 - |\max_E|) + \cdots + (x_{n-1} - |\max_E|) \geq k - (n-1) \times |\max_E|$$

$$\frac{x_1-|\max_E|}{|\min_E-|\max_E||} + \cdots + \frac{x_{n-1}-|\max_E|}{|\min_E-|\max_E||} \geq \frac{k-(n-1)\times|\max_E|}{|\min_E-|\max_E||}$$

$$2^{\frac{x_1-|\max_E|}{|\min_E-|\max_E||}} \times \ldots \times 2^{\frac{x_{n-1}-|\max_E|}{|\min_E-|\max_E||}} \geq 2^{\frac{k-(n-1)\times|\max_E|}{|\min_E-|\max_E||}}$$

that is the wining probability in a fair tree run for ω^* in G' □

5 Heuristics and Experimental Evaluation

Our hardness results lead us to conjuncture that the general problem of rigging an election agenda with incomplete information is NP-Hard to compute; but we also think that a worst-case analysis is not enough. The situation is similar to that in cryptography, where a secure protocol is not one that is hard to break in the worst case, but one that can be broken only with negligible probability. Bartholdi et al., who were in many ways pioneers of the complexity-theoretic approach to understanding election manipulation [3] first voiced the concern that NP-Hardness results are not enough:

Concern: It might be that there are effective heuristics to manipulate an election even though manipulation is NP-complete.
Discussion: True. The existence of effective heuristics would weaken any practical import of our idea. It would be very interesting to find such heuristics.

This motivates us to consider heuristics for this problem, and to test their performance in different scenarios. We used a simulation program (written in C) to evaluate the heuristics. Our experiments are in two categories: first, randomly generated data, and second, public domain form data from sports competitions. For each of these settings, we evaluated the heuristics for both fair trees and linear orders. For the randomly generated data sets, we first generated random values for the probability matrix from a uniform distribution in the range $[0, 1]$, and completed the matrix to preserve probability constraints. We then ran 100 iterations, and during each iteration a winner candidate, ω^*, was randomly chosen and each heuristic was used in an attempt to generate an optimal order for this candidate. The second scenario was the same, except that the random values for the matrix were taken from a normal distribution with an average of 0.5 and a standard deviation of 0.2. (If a value of more than 1 (less than 0) was generated it was changed to 1 (0) to preserve probability constraints.) For the real-world data sets, we based our experiments on data from basketball and tennis competitions. For the basketball experiments, we took the 29 teams in the NBA, and computed the ballot matrix from public domain form data. Here, there were no iterations, but for every team we used each heuristic to generate a playing order that would give this team the best chance of winning. For the tennis experiments, we used the 13 players at the top of the ATP ranking, again computing the ballot matrix from public domain form tables.

5.1 Heuristics for Linear Voting Orders

The heuristics we developed are as follows.

- *Optimal*: In those cases where it was computationally feasible to do so, we exhaustively evaluated every permutation in order to find the real optimal solution as a comparison.
- *Far adversary*: The idea here is to put candidates who are likely to beat you as far away from you as possible. Thus, the candidate who has the highest probability to beat ω^* was selected to be the first, and so on; ω^* was chosen to be the last (cf. Corollary 1).
- *Best win*: ω^* was chosen to be the last, and the candidate that ω^* has the best chance of beating was chosen to be before it, the next being the one that this candidate was most likely to beat, and so on.
- *Simple convert*: The idea here is to convert the imperfect information ballot matrix into one of perfect information, and then simply apply an algorithm which is known to work in polynomial time for such cases [8]. To create the perfect information matrix, every probability which was greater than or equal to 0.5 was converted to 1, and others were converted to 0. If no order could be found, (which is sometimes the case where the number of candidates is small) a random order was generated.
- *Threshold convert*: This is a more sophisticated attempt to make a perfect information ballot matrix. We searched for the maximum threshold above which, if we convert all the probabilities above this to 1 and below it to 0, we still have an order that enable ω^* to win (on the converted tournament matrix). We used a binary search, stopping when the difference between the low/high limit and the threshold was less than 0.005. As before, if no order could be found a random order was generated.

- *Local search*: ω^* was chosen to be last. For the other places, in every iteration a random permutation was chosen. Then $0.5 * num\text{-}of\text{-}candidates$ random swaps were tested to find an order with maximum winning probability. ($num\text{-}of\text{-}candidates$ iterations were done.)
- *Random order*: As a control, a random order was also generated.

The results for heuristics on linear orders are shown in Figure 2.

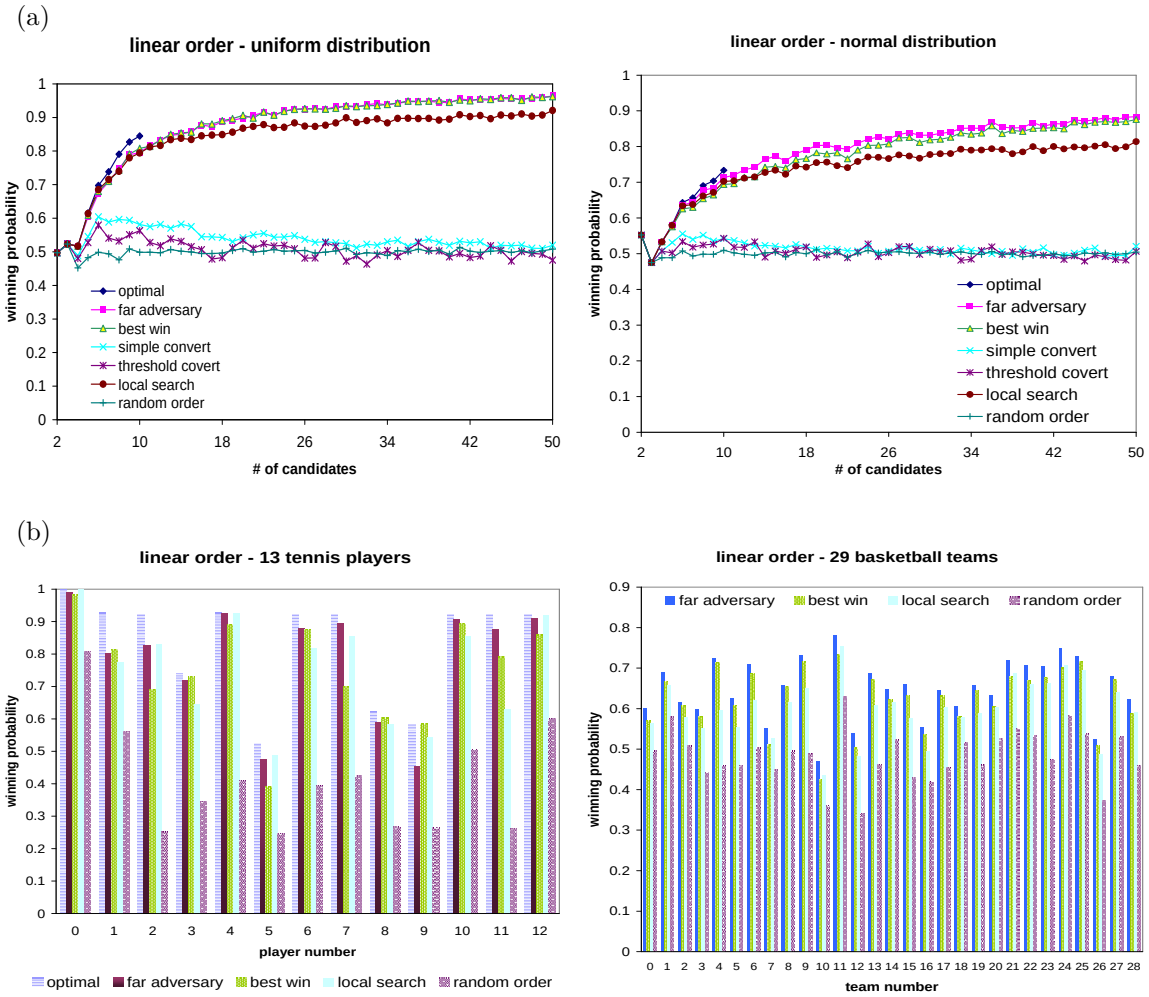


Figure 2: Performance of heuristics for linear order ballots. (a) Shows the performance of the heuristics for randomly generated ballot matrices using a uniform and normal probability distributions; and (b) Shows performance for real-world data from the domain of professional tennis and basketball.

First, note that the overall performance of the heuristics does not vary significantly between uniform and normal distributions (Figure 2(a) left and right columns, respectively). In these graphs, the x-axis is the number of candidates and the y-axis is the winning probability that was found using our heuristics. Every point in the graph represents the winning probability that was averaged over 100 iterations.

In the uniform distribution experiments, it seems that best-win and far adversary seem to perform similarly well, (marginally better than local search), while for a normal distribution they are slightly differentiated, but again perform better than local search. In the graph for linear order with normal distribution (Figure 2(a) left column), again, we find the first two heuristics – far adversary and best win – gave about the same results, while local search lies a little behind; all reach a very high winning probability, almost 0.9, and they performed better as the number of candidates increased. These heuristics also perform well when comparing them to the optimal solution – they gave a winning probability which is on average only 98% from the optimal solution. Note that the "convert" heuristics perform very poorly, both for uniform and normal distributions, and so we omitted them in subsequent experiments.

In the graph for linear order with 13 tennis players (Figure 2(b), left column), the x-axis is the player number that was chosen to be the winning candidate and the y-axis as before. Here, there was no heuristic that performed significantly better than the others in general, but when choosing from the heuristics, the best solution for each candidate performs very well when compared to the optimal solution. They gave a winning probability which is only 96% from the optimal solution on average. The winning probability is on average more than twice as high as the random order. Player number 0, (Roger Federer, currently the world's number one player), even succeeded in obtaining a winning probability of 1 from local search.

We conclude that there is no one heuristic that performs significantly better than the others for all cases. We suggest the best thing to do here is to run all the heuristics and order the candidates according the heuristic which gives the best results for this candidate because they all run quite fast. Note that local search takes much more time than other heuristics, but still has acceptable time performance.

5.2 Heuristics for Fair Voting Trees

For this voting agenda type we investigated the following heuristics.

- *Optimal*, *Far adversary*, *Local search*: We organized the leaves of the binary tree as a linear order from left to right and applied these heuristics as above. (Note that when the number of candidates is not a power of 2, some of the rightmost candidates may face one less ballot than other candidates.)

- *Best win*: Because of the tree structure, we had to use a modified version of the best win heuristic, but the principle remains the same. We try to maximize the probability that ω^* will face candidates that he has a high probability of beating, and to maximize the probability that they will reach the point when they compete against him. So we first assign ω^* at the rightmost leaf of the voting tree, and for each ballot along its path to the root we assign candidates that ω^* has a high probability to beat. In this way we define for each one of them a sub-tree that we want him to be its overall winner (unless this candidate has been assigned to a leaf) so we can repeat this assignment procedure recursively.

The results of our experiments with balanced tree ballots are shown in Figure 3.

In the tree order with random uniform and normal distribution (Figure 3(a) left and right columns, respectively) the axes are as in the linear order case. Generally, the wining probability is much lower than the linear order voting protocol, which seems a direct consequence of the relative fairness of the procedure. Nevertheless, if we compare the best heuristic for each case to the random order, we get a winning probability which is on average 4.31 times higher than the random order winning probability and which is on average only 96% from

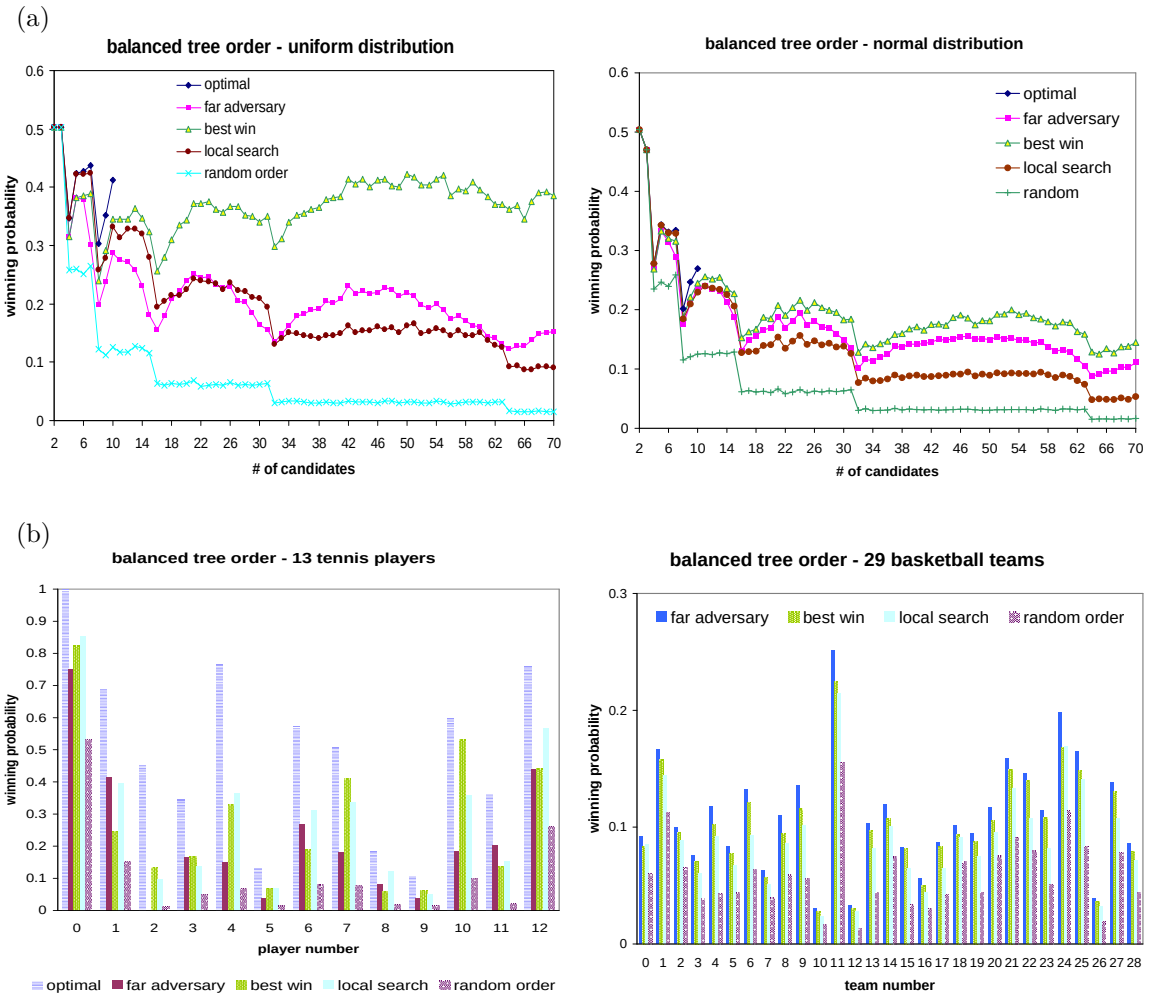


Figure 3: Performance of heuristics for balanced tree ballots. (a) Shows the performance of the heuristics for randomly generated ballot matrices using a uniform and normal probability distributions; and (b) Shows performance for real-world data from the domain of professional tennis and basketball.

the optimal solution. We also note that the graphs have a wave-like shape: the lowest points in these waves are when the number of candidates is exactly a power of 2. This is simply because when the number of candidates is a power of 2, every candidate must compete in exactly the same number of ballots. In all other cases, there are some of the candidates (including our favorite, ω^*) which face one less ballot. This effect can help ω^*, by forcing strong competitors to undergo one more ballot. It can also be seen that in contrast to the linear order with random normal distribution, the best win heuristic performs better than the others: 1.25 times better than far adversary, and 1.66 times better than local search on average.

In the tree order with 29 NBA basketball teams graph, (Figure 3(b), right column), the far adversary method performed better than others with a winning probability that is on average 1.24 times better than local search and 1.08 times better than best win. The highest winning probability was generated for team 11, the LA Lakers. It was not computationally feasible to calculate the optimal solution for 29 teams, so we ran another scenario with only the first 13 teams to check the performance of our heuristics against the optimal solution. The best heuristic in each case gave a winning probability which in on average only 99% from the optimal solution!

In the tree order with 13 tennis players graph ((Figure 3(b), left column) there was no heuristic that performed significantly better than the others, just as in the linear order case. But here, when we choose from the heuristics the best solution for each case they gave a winning probability which is on average 61% from the optimal solution. Perhaps the performance difference between this case and the basketball case emerges from the type of the distribution. The basketball scenario has a distribution which is much closer to a normal distribution than in the tennis scenario, in which the probability matrix contains many high probabilities. Another indicator for this is the performance of our best heuristic in each case against the random order. In the tennis scenario it was almost 5 times better, while in the basketball scenario it was only about 1.5 times better.

6 Related Work and Conclusions

A closely related stream of work is the problem of optimal seeding for tournaments. This problem considers how to determine an agenda for a voting tree that will result in an "interesting" sporting competition. [10] for example, assumes an imperfect information ballot matrix as we do, and uses it to produce an interesting competition agenda that still satisfies some fairness criterions. [7] investigates a 4-player scenario with an auction-like analysis; instead of knowing the probability of winning, each player exerts some effort according to its private valuation. Early work by [11] analyzes voting trees for 8 players, among 3 other tournament types. It uses an imperfect information matrix and investigates the effect of the initial agenda on the probability that the best player will win the game. Note that none of these papers analyzed the complexity of finding the optimal agenda for a specific candidate, however.

By understanding when the manipulation of elections is possible, and those cases where it is computationally easy to manipulate an election, we can engineer voting protocols to avoid such cases. We have demonstrated that, while it seems hard to control an election under incomplete information in theory, there are heuristics that perform well on this problem in practice. This research can hence usefully inform the design of future voting protocols in multi-agent systems.

References

[1] K. J. Arrow, A. K. Sen, and K. Suzumura, eds. *Handbook of Social Choice and Welfare Volume 1.* Elsevier, 2002.

[2] J. Bang-Jensen and G. Gutin. On the complexity of hamiltonian path and cycle problems in certain classes of digraphs. *Discrete App. Math.*, 95:41–60, 1999.

[3] J. J. Bartholdi, C. A. Tovey, and M. A. Trick. The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6:227–241, 1989.

[4] J. J. Bartholdi, C. A. Tovey, and M. A. Trick. How hard is it to control an election. *Math. and Comp. Modelling*, 16:27–40, 1992.

[5] S. J. Brams and P. C. Fishburn. Voting procedures. In K. J. Arrow, A. K. Sen, and K. Suzumura, eds, *Handbook of Social Choice and Welfare Volume 1*, Elsevier, 2002.

[6] U. Endriss and J. Lang, editors. *Proc. COMSOC-2006.* Amsterdam, 2006.

[7] Christian Groh, Benny Moldovanu, Aner Sela, and Uwe Sunde. Optimal seedlings in elimination tournaments. *Jnl of Economic Th.*, 107:140–150, 2008.

[8] J. Lang, M. S. Pini, K. B. Venable, and T. Walsh. Winner determination in sequential majority voting. In *Proc. IJCAI-07*, 2007.

[9] T. Sandholm. Distributed rational decision making. In G. Weiß, ed., *Multiagent Systems*, pages 201–258. MIT Press 1999.

[10] Allen J. Schwenk. What is the correct way to seed a knockout tournament. *American Math. Monthly*, 107(2):140–150, 2000.

[11] Donald T. Searls. On the probability of winning with different tournament procedures. *Jnl. of the American Statistical Assn*, 58(304):1064–1081, 1963.

Noam Hazon
Department of Computer Science, Bar Ilan University
Ramat Gat 52900, Israel
Email: `hazonn@cs.biu.ac.il`

Paul E. Dunne
Department of Computer Science, University of Liverpool
Liverpool L69 7ZF, United Kingdom
Email: `ped@liverpool.ac.uk`

Sarit Kraus
Department of Computer Science, Bar Ilan University
Ramat Gat 52900, Israel
Email: `sarit@cs.biu.ac.il`

Michael Wooldridge
Department of Computer Science, University of Liverpool
Liverpool L69 3BX, United Kingdom
Email: `mjw@liverpool.ac.uk`

A geometric approach to judgment aggregation

Christian Klamler and Daniel Eckert

Abstract

Don Saari has developed a geometric approach to the analysis of paradoxes of preference aggregation such as the Condorcet paradox or Arrow's general possibility theorem. In this paper we extend this approach to judgment aggregation. In particular we use Saari's representation cubes to provide a geometric representation of profiles and majority rule outcomes. We then show how profile decompositions can be used to derive restrictions on profiles that guarantee logically consistent majority outcomes. Moreover, we use our framework to determine the likelihood of inconsistencies. Finally, current distance-based approaches in judgment aggregation are discussed within our framework.

1 Introduction

The problem of judgment aggregation consists in aggregating individual judgments on an agenda of logically interconnected propositions into a collective set of judgments on these propositions. This relatively new literature (see List and Puppe [6] for a survey) is centred on problems like the discursive dilemma which are structurally similar to paradoxes and problems in social choice theory like the Condorcet paradox and Arrow's general possibility theorem. For the analysis of such paradoxes Saari [9] has successfully introduced a geometric approach, the extension of which to judgment aggregation seems promising.

A major difference of judgment aggregation to social choice theory lies in the representation of the information involved. While binary relations over a set of alternatives are a canonical representation of preferences, a natural representation of judgments are binary valuations over a set of propositions, where the logical interconnections between these propositions determine the set of admissible valuations. E.g. the agenda of the famous discursive dilemma $\{p, q, p \wedge q\}$ is associated the set of admissible, i.e. logically consistent valuations $\{(0,0,0),(1,0,0),(0,1,0),(1,1,1)\}$.

In this paper we want to use Saari's tools to derive and analyse results in judgment aggregation. In section 2 we introduce the formal framework. Section 3 uses Saari's representation cubes to provide a unified geometric representation of profiles and majority rule outcomes. This will clarify which problems can occur in judgment aggregation using majority rule on certain domains. Applying Saari's idea of a profile decomposition, we also show how majority inconsistencies can be avoided with the help of restrictions on the distribution of individual valuations, i.e. a kind of generalized domain restriction. This leads us to the determination of the likelihood of inconsistencies under majority rule for different agendas in section 4. In section 5, we apply our approach to illuminate current results on distance-based judgment aggregation. Finally, section 6 concludes the paper.

2 Formal Framework

Let J be the set of propositions on which judgments have to be made. Most problems in the literature on judgment aggregation can be formulated with the help of vectors of binary valuations $x = (x_1, x_2, ..., x_{|J|}) \in X \subseteq \{0,1\}^{|J|}$, where $x^j = 1$ means that proposition j is

believed and X denotes the set of all admissible (logically consistent) valuations (see Dokow and Holzman [2]).

A profile of individual valuations is represented by a vector $\mathbf{p} = (p_1, p_2, ..., p_{|X|})$ which associates with every binary valuation $x_k \in X$ the fraction p_k of individuals with this valuation. This is an anonymous representation of voters' valuations as only the distribution of the valuations matters.

A judgment aggregation rule is a mapping f that associates with every profile $\mathbf{p} = (p_1, p_2, ..., p_{|X|})$ a valuation $f(\mathbf{p}) \in \{0,1\}^{|J|}$. If $f(\mathbf{p}) \in X$ for all $\mathbf{p}$ we will call its domain f-consistent.

Saari [9] analyses preference aggregation using a geometric approach. For the simplest setting consider three alternatives a, b, c. This gives rise to three issues, i.e. pairwise comparisons, namely between a and b, b and c and c and a. A "1" for the first issue (i.e. the comparison between a and b) means that a is preferred to b, written $a \succ b$. On the other hand, a "0" indicates the opposite preference, i.e. $b \succ a$. Defining the average support for issue j by $x_j = \frac{\sum_{i \in N} x_j^i}{n}$, where N denotes the set of individuals, a preference profile maps into a point $x \in [0,1]^{|J|}$ in the hypercube with dimension $|J| = 3$ (the number of issues, i.e. pairwise comparisons). See figure 1.

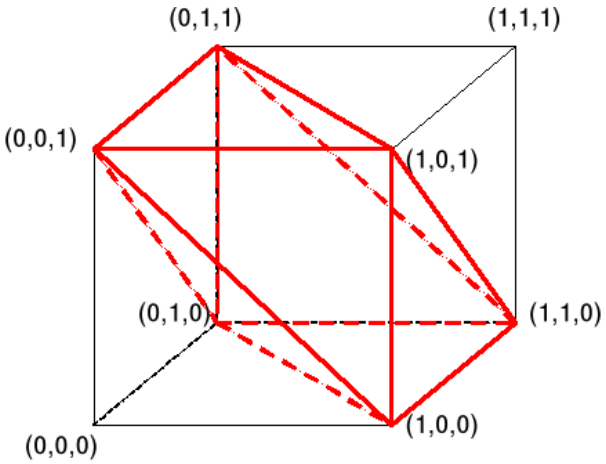


Figure 1: Saari's representation cube

In figure 1, the vertex $(0,0,1)$ thus represents the preference where $b \succ a$, $c \succ b$ and $c \succ a$ or - for simplicity - the ranking *cba*. As there are eight vertices but only six transitive rankings of the three alternatives, there are two vertices representing irrational voters with cyclic preferences, namely $(0,0,0)$ and $(1,1,1)$. If we exclude those vertices, we see that the convex hull of the remaining six vertices is the representation polytope, i.e. every preference profile maps into a point in this polytope.

3 Majority (In)consistency and Domain Restrictions

The same 3-dimensional hypercube can be used for a simple judgment aggregation problem with 3 propositions (issues), i.e. $|J| = 3$. For simple majority voting on the issues, every profile $\mathbf{p}$ of individual judgments on J is mapped into a point $x(\mathbf{p})$ in the hypercube. Its Euclidean distance to the respective vertices determines the majority outcome. This means that the hypercube can be partitioned into 8 equally sized subcubes each determining the majority outcome for profiles mapped into those subcubes. E.g. in figure 2 the shaded

subcube, determined by the diagonal $\left[(\frac{1}{2},\frac{1}{2},\frac{1}{2}),(1,0,1)\right]$ consists of all points that are of closest Euclidean distance to the vertex $(1,0,1)$ and hence any $x(\mathbf{p})$ in that subcube leads to a majority outcome of $(1,0,1)$. For $d_E(x,y)$ denoting the Euclidean distance between $x,y \in \{0,1\}^{|J|}$, we can also think of the majority valuation x^M as the

$$argmin_{x\in\{0,1\}^{|J|}} \sum_{k=1}^{|X|} p_k d_E(x^k,x)$$

.

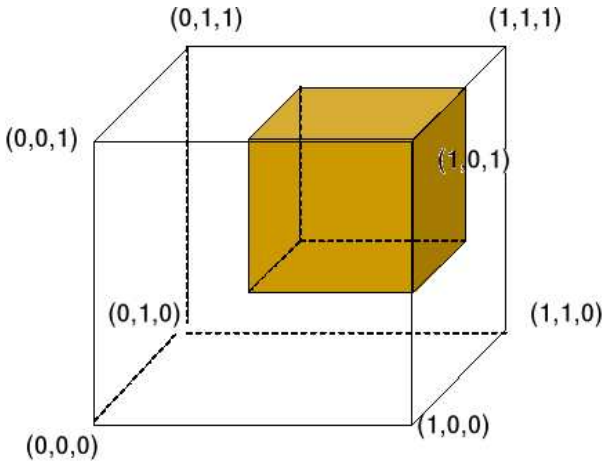


Figure 2: Majority subcube

Consider the agenda $\{p, q, p \wedge q\}$ of the discursive dilemma with associated domain of admissible valuations $X = \{(0,0,0),(1,0,0),(0,1,0),(1,1,1)\}$. The four admissible vertices in the hypercube determine the representation polytope as seen in figure 3.

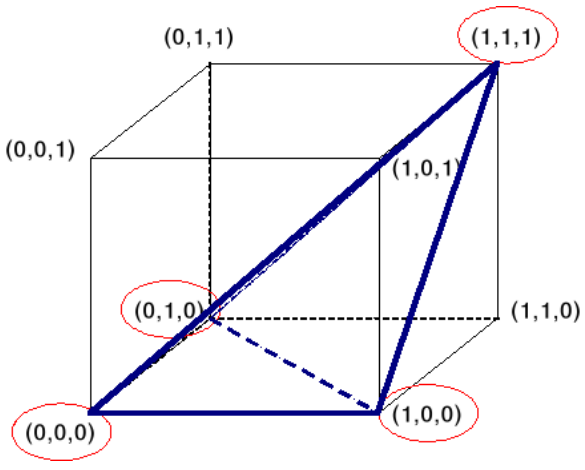


Figure 3: representation polytope

Given X, consider the profile $\mathbf{p} = (0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, i.e. no voter has valuation $(0,0,0)$, one third of the voters has valuation $(1,0,0)$, and so on. As this maps into the point $x = (\frac{2}{3}, \frac{2}{3}, \frac{1}{3})$ - a point whose closest vertex is $(1,1,0)$ - the representation polytope obviously passes

through one subcube representing a majority outcome not in the domain. That such an inconsistency can easily occur in general is seen from the following lemma:

Lemma 1 *Given any vertex $x \in \{0,1\}^{|J|}$, there exist 3 vertices a, b, c such that for some linear combinations of those vertices there is a point in the x-subcube.*

For $|J| = 3$, these 3 vertices necessarily need to be the neigbors of that vertex, i.e. they are only allowed to differ from it in one issue. Given that, we can now provide a simple result for the occurrence of majority consistency, i.e. what X needs to look like to guarantee that the majority outcomes are themselves in X.

Proposition 1 *For $|J| = 3$, the set of valuations X is majority consistent iff for any triple of vertices in the domain with a common neighbor, this common neighbor is also contained in the domain.*

To analyse those paradoxical outcomes and suggest restrictions to overcome those, we will use a profile decomposition technique developed by Saari [9]. Consider two indviduals with the respective valations $(1,0,0)$ and $(0,1,1)$. They are exact opposites, so from a majority rule point of view those two valuations cancel out. Hence this implies that for any two opposite admissible valuations in X, we can cancel the valuation held by the smaller number of individuals. This leads to a reduced profile, the majority outcome of which is identical to the majority outcome of the original profile.

E.g. given $X = \{(0,0,0),(1,0,0),(0,1,0),(1,1,1)\}\}$, with (p_1, p_2, p_3, p_4) being the shares of individuals holding each of the respective valuations where $\sum_i p_i = 1$. As $(0,0,0)$ and $(1,1,1)$ are exact opposites, the reduced profile will have a share of 0 for the valuation held by the smaller number of individuals. In the case of $p_1 > p_4$ such a reduced profile will be $\mathbf{p}' = (\frac{p_1 - p_4}{p_1+p_2+p_3-p_4}, \frac{p_2}{p_1+p_2+p_3-p_4}, \frac{p_3}{p_1+p_2+p_3-p_4}, 0)$, in the case of $p_1 \leq p_4$ we can create the reduced profile accordingly. Hence the reduced profile maps into one of the following two planes represented in figure 4, namely either into the one determined by the vertices $(0,0,0)$, $(1,0,0)$ and $(0,1,0)$ or the one determined by the vertices $(1,0,0)$, $(0,1,0)$ and $(1,1,1)$.

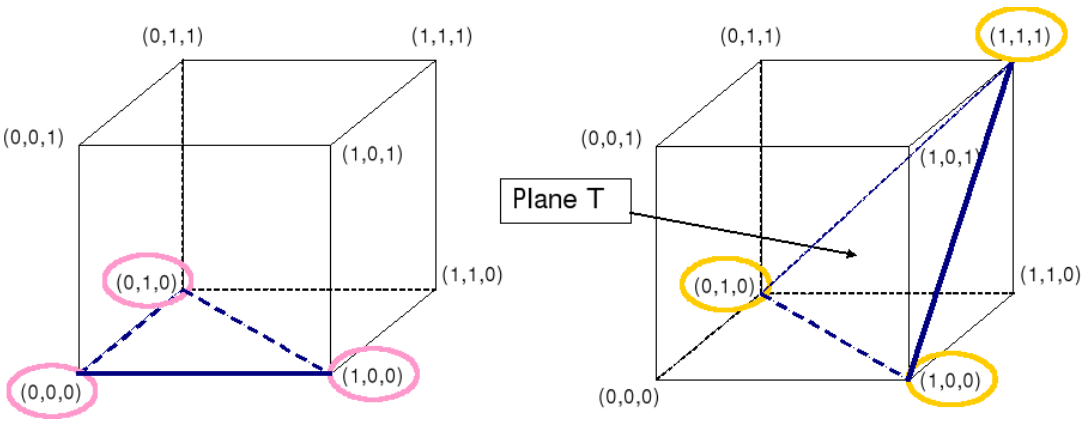


Figure 4: planes

Now we want to determine whether there are consistency conditions, i.e. what sort of profiles do guarantee majority consistency in the sense of a majority outcome being one of the valuations in X. Let $a_i = \frac{p_i}{\sum_{i=2}^{4} p_i}$, for $i = 2,3,4$. Then $\alpha = (a_2 + a_4, a_3 + a_4, a_4) \in T$. By definition, $\mathbf{p} = (1 - p_1)(0, a_2, a_3, a_4) + p_1(1,0,0,0)$. By linearity, $x(\mathbf{p}) = (1 - p_1)\alpha +$

$p_1(0,0,0) = (1-p_1)\alpha$, where $x(\mathbf{p}) \in [0,1]^3$ is a vector summarizing the average support for each issue.

So, geometrically any profile can be plotted via a point in the plane T, its connection to the $(0,0,0)$ vertex and a weight p_1. The following figure 5 shows plane T, the shaded area of which indicates the cut with the $(1,1,0)$-subcube and hence those points where a profile leads to an inadmissible majority outcome.

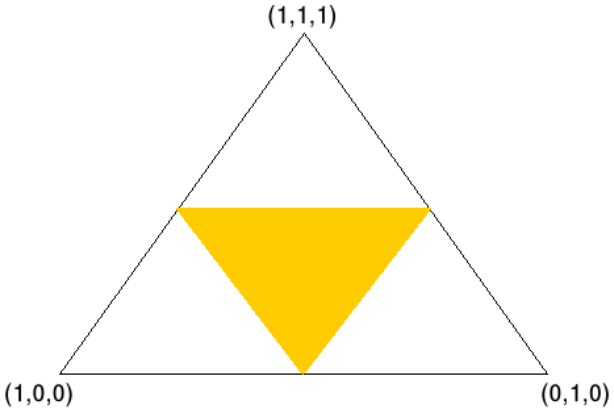


Figure 5: plane T

Now, we can state the following proposition:

Proposition 2 *Given* $X = \{(0,0,0),(1,0,0),(0,1,0),(1,1,1)\}\}$ *and* $\mathbf{p} = (p_1,p_2,p_3,p_4)$ *with* $\sum_{i=1}^{4} = 1$*, satisfying one of the following restrictions on the distribution of* $\mathbf{p}$ *is necessary and sufficient for admissible majority outcomes:*

(a) $p_1 \geq p_4$

(b) $\frac{p_2}{\sum_{i=2}^{4} p_i} > \frac{1}{2}$

(c) $\frac{p_3}{\sum_{i=2}^{4} p_i} > \frac{1}{2}$

(d) $\frac{p_4}{\sum_{i=2}^{4} p_i} > \frac{1}{2}$

Proof. If $p_1 \geq p_4$, the reduced profile is mapped into the plane indicated on the left of figure 4 which is closed under majority rule. Otherwise $p_1 < p_4$ and the reduced profile is mapped into plane T. As we see on the right side of figure 4, majority consistency is guaranteed in the white triangles, and a profile maps into one of them whenever one of restrictions $(b) - (d)$ is satisfied. On the other hand, given that for $|X| = 4$ and $|J| = 3$, elementary algebra shows that every profile maps into one and only one point in the representation polytope. Hence, majority consistency is satisfied only if one of the above restrictions is fulfilled. ■

One interesting feature of those restrictions is that they are based on the space of profiles which is more general than restrictions on the space of valuations which is usually used in classical domain restrictions. E.g. List [5] introduces the unidimensional alignment domain which has a certain resemblence to Black's single peakedness condition in social choice theory. It requires individuals to be ordered from left to right such that on each issue there occurs only one switch from acceptance to non-acceptance (or vice versa). For $|J| = 3$ a

unidimensional alignment domain does satisfy one of the above conditions for admissible majority outcomes.[1]

Moreover, this framework also opens the analysis of various paradoxical situations, e.g. strong support for one particular issue but still inadmissible majority outcomes. This is stated in the following proposition:

Proposition 3 *There exist profiles such that there is almost unanimous agreement on one issue and still an inadmissible majority outcome is obtained.*

Proof. Looking at figure 5 one observes, that points close to the edge connecting the vertices $(1,0,0)$ and $(1,1,1)$ have almost unanimous agreement on issue 1. However, at the midpoint of this edge, the shaded triangle comes arbitrarily close to the edge. Hence, there exist profiles which lie in the shaded triangle but imply almost unanimous agreement on one issue. The same argument applies to points close to the edge conncecting the vertices $(0,1,0)$ and $(1,1,1)$. ■

4 Likelihood of Inconsistency

The geometric framework can also be used to analyze the likelihood of inadmissible outcomes. The approach is based on the fact that only 4 vertices are admissible individual valuations, and hence any majority outcome in the representation cube is determined by a unique profile. Consider again the situation $X = \{(0,0,0),(1,0,0),(0,1,0),(1,1,1)\}\}$. Then for any vector of shares of individual valuations $\mathbf{p} = (p_1, p_2, p_3, p_4)$ we get the following for the average support on each issue: $x_1 = p_2 + p_4$, $x_2 = p_3 + p_4$, $x_3 = p_4$, $1 = p_1 + p_2 + p_3 + p_4$. As those are 4 equations with 4 unknowns there exists a unique solution. Thus, assuming every profile being equally likely - i.e. taking an impartial anonymous culture[2] - the volume of certain subspaces now indiciates the likelihood of occurrence of certain outcomes. Consider first the volume of the representation cube: $V = \frac{1}{2} \cdot 1 \cdot \frac{1}{3} = \frac{1}{6}$. On the other hand, points leading to inadmissible majority outcomes are located in the tetraeder determined by the points $\left[(\frac{1}{2},\frac{1}{2},0),(1,\frac{1}{2},\frac{1}{2}),(\frac{1}{2},1,\frac{1}{2}),(\frac{1}{2},\frac{1}{2},\frac{1}{2})\right]$. The volume of this tetraeder is $\frac{1}{48}$ (see figure 6).

So its volume relative to the volume of the whole representation polytope is $\frac{1}{8}$ and hence we can say that the probability of an outcome being inadmissible is 12,5 percent. This provides a different approach to derive the expected probability of paradoxical situations under the impartial anonymous culture compared to the non-geometric approach by List [5], leading to the same results.

Of course, different domains allow for different probabilities. E.g. consider the agenda $\{p, q, p \leftrightarrow q\}$ with $X = \{(0,0,1),(1,0,0),(0,1,0),(1,1,1)\}\}$. Then, for any point $x = (x_1, x_2, x_3)$ in the representation polytope we get $x_1 = p_2 + p_4$, $x_2 = p_3 + p_4$, $x_3 = p_1 + p_4$ and $p_1 + p_2 + p_3 + p_4 = 1$. Again, every profile maps into a unique point in the representation polytope. Making the same volume calculations as before, we get - under the impartial anonymous culture - a probability of inadmissible outcomes of 25 percent.

5 Codomain Restrictions and Distance-Based Aggregation

Besides restrictions on the space of profiles, there is an alternative way to guarantee logical consistency at the collective level, namely via restricting the set of collective outcomes

[1] For a more elaborated discussion on majority voting on restricted domains see also Dietrich and List [1].

[2] See Gehrlein [3] for a general discussion of the impartial anonymous culture.

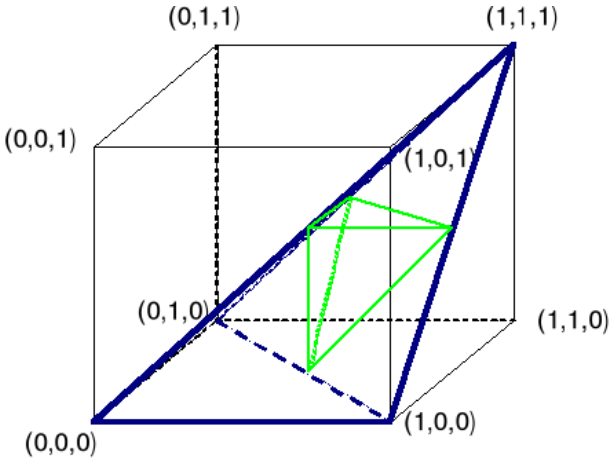


Figure 6: Distances

to admissible valuations. One way to work with such codomain restrictions is by using distance-based aggregation rules. In analogy to a well-known procedure in social choice theory (Kemeny [4]), Pigozzi [8] introduced such an approach to judgment aggregation. In principle a distance-based aggregation rule determines the collective valuation as the valuation that minimizes the sum of distances to the individual valuations. Formally, given the profile of individual valuations $(x^1, x^2, ..., x^n)$, the collective valuation is the admissible valuation $x \in X$ that minimizes the sum of distances to the individual valuations, i.e.

$$f(\mathbf{p}) = argmin_{x \in X} \sum_{i=1}^{n} d(x, x^i)$$

The most commonly used distance function is the Hamming distance, which counts the number of issues on which two valuations disagree, i.e. for $x = (1, 0, 0)$ and $x' = (1, 1, 1)$, $d(x, x') = 2$.

It is easily seen that on a majority consistent domain this distance-based aggregation rule coincides with majority voting on issues, thus providing a metric rationalization of majority voting.

Now given our geometric approach, there is a simple geometric explanation of this distance-based aggregation rule. As could be seen in figure 5, all problematic profiles lead to a point in the shaded triangle. However, one option is to divide the triangle into three sub-triangles as in figure 7.

The point in the middle is exactly the barycenter point of the triangle $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Using these additional lines, we now have divided the triangle into three areas, points in which are characterized by being of smallest Euclidean distance to the vertex of the proper triangle w.r.t. the points within the shaded triangle. So points in the south-western part of the shaded triangle will be closest to the $(1, 0, 0)$ vertex. This, however, is identical to saying that for any point in the shaded triangle, switch the majority valuation on the issue which is closest to the 50-50 threshold (see Merlin and Saari [7]).

Example 1 *Let* $X = \{(0, 0, 0), (1, 0, 0), (0, 1, 0), (1, 1, 1)\}$ *and* $\mathbf{p} = (0.1, 0.35, 0.3, 0.25)$. *This leads to* $x(\mathbf{p}) = (0.6, 0.55, 0.25)$ *and hence an inadmissible majority outcome* $(1, 1, 0)$. *Looking at figure 5 we see that* $\alpha \in T$ *lies in the south-western shaded triangle. Thus, according to our distance-based aggregation rule, the outcome will be the admissible valuation* $(1, 0, 0)$

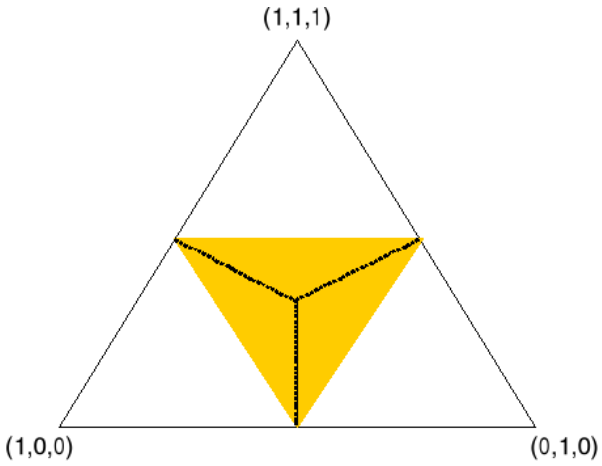


Figure 7: Distances

as α is closest to the $(1,0,0)$ vertex. However, this can also be seen as switching the valuation on the issue which is closest to the 50-50 threshold, which - in $x(\mathbf{p})$ - is obviously issue 2.

6 Conclusion

In this paper we have shown how geometry can be used to analyse results in judgment aggregation, such as majority inconsistencies and distance based aggregation rules, and to determine new results such as guaranteeing majority consistency via restrictions on distributions of individual valuations and determining the likelihood of inconsistencies.

Most of the stated results do not easily extend to more than three issues because of problems of dimensionality. E.g. an agenda with three propositions and their conjunction, like $\{p, q, r, p \wedge q \wedge r\}$, leads to eight admissible valuations, i.e. eight vertices out of the 16 vertices in the four-dimensional hypercube. The extensions of our (domain) restrictions and calculations of the likelihood of the occurrence of paradoxes to those higher dimensions are not obvious and need further work.

References

[1] Dietrich, F. and C. List (2007): Majority voting on restricted domains. *mimeo.*

[2] Dokow, E. and Holzman, R. (2005), Aggregation of binary evaluations, *mimeo.*

[3] Gehrlein, W.V. (2006): *Condorcet's Paradox.* Springer, Berlin.

[4] Kemeny, J. (1959): Mathematics without numbers. *Daedalus* 88, 4, 577–591.

[5] List, C. (2005): The probability of inconsistencies in complex collective decisions. *Social Choice and Welfare* 24, 3–32.

[6] List, C. and C. Puppe (2007): Judgment aggregation: a survey. mimeo.

[7] Merlin, V. and D.G. Saari (2000): A geometric examination of Kemeny's rule. *Social Choice and Welfare* 17, 3, 403–438.

[8] Pigozzi, G. (2006): Belief merging and the discursive dilemma: an argument-based account to paradoxes of judgment aggregation. *Synthese* 152, 2, 285–298.

[9] Saari, D.G. (1995): Basic geometry of voting. Springer, Berlin.

Christian Klamler, Daniel Eckert
Institute of Public Economics
University of Graz
8010 Graz, Austria
Email: {christian.klamler,daniel.eckert}@uni-graz.at

Judgment Aggregation as Maximization of Epistemic and Social Utility

Szymon Klarman

Abstract

We restate the problem of judgment aggregation and approach it using the decision-theoretic framework applied by I. Levi to modeling acts of rational acceptance in science. We propose a method of aggregation built on the concepts of epistemic and social utility of accepting a collective judgment, which accounts for such parameters as the factual truth of the propositions, reliability of agents, information content (completeness) of possible collective judgments and the level of agreement between the agents. We argue that the expected utility of accepting a judgment depends on the degree to which all those objectives are satisfied and that groups of rational agents aim at maximizing it while solving judgment aggregation problems.

1 Introduction

The problem of judgment aggregation, concerning the issue of building up collective judgments by groups of agents, lies at the intersection of social choice theory and epistemology. On the one hand it deals with the question of what is a good procedure by which individual viewpoints should contribute to the collective one — a central matter of concern of social choice theory. On the other, since the objects between which the choice is made are *judgments*, any proposed method of aggregation has to be verified against logical, or even broader, epistemological criteria, guaranteeing soundness of the outcomes. For that reason the problem falls also in the scope of some essentially philosophical considerations, of which the most important regards the rationality of acceptance of propositions in general.

Due to the twofold character of the problem an aggregation procedure is expected to be socially fair and epistemologically reliable at the same time. It might be the case, as C. List and P. Pettit show in [12], that these two requirements cannot be satisfied simultaneously and one has to prioritize between them. Nonetheless, even if this state of affairs is inevitable, it is still worth asking where exactly the trade-off takes place and whether it is possible to capture it formally and gain substantial control over it.

An interesting framework for such an analysis has been proposed by I. Levi [9, 10], who applied the notion and a simple measure of *epistemic utility* of accepting a proposition in order to deal with the problem of underdetermination of inductive inferences in science.[1] As a result, induction has been formally reinterpreted in terms of the trade-off between rival epistemic goals that drive scientific inquiry. This approach has been recently recalled in the context of judgment aggregation by Levi himself [11] and also by D. Fallis in [3]. In this paper we propose and discuss a possible aggregation procedure based on those grounds, which explicitly parameterizes the trade-offs underlying the problem of aggregating judgments, accounting for the relevant epistemological and social-theoretical aspects.

The reminder of the paper is organized as follows. First, we outline the formal frames for the judgment aggregation problem together with the discursive dilemma. Then we point out a correspondence between the dilemma and the lottery paradox and recall the work of I. Levi used for circumventing the latter. In Section 3 we introduce the utilitarian judgment aggregation model, followed by sample aggregation results and the discussion of the method.

[1] The idea was first brought about by R. Jeffrey [6] and C. G. Hempel [5], however the approach was not formally elaborated and introduced as a self-standing proposal until Levi's work.

2 Discursive Dilemma and Lottery Paradox

The *judgment aggregation problem* can be shortly characterized as follows (cf. [12, 1]). Let $\mathcal{A} = \{1, \ldots, n\}$ be a set of n agents and Φ an *agenda*, i.e. a set of well-formed propositional formulas. It is assumed that for every $\varphi \in \Phi$ there is also $\neg\varphi \in \Phi$. Each agent has an *individual set of judgments* with respect to Φ. A judgment regarding a proposition is understood here as an unequivocal act of *acceptance* or *rejection* of that proposition. An individual set of judgments of agent i can be represented as a subset $\Phi_i \subseteq \Phi$ of exactly those propositions that are accepted by i. All the propositions that are not in Φ_i are the ones that i rejects. Further, we require every individual set of judgments Φ_i to satisfy three rationality constraints: 1) *completeness*: for every $\varphi \in \Phi$, either $\varphi \in \Phi_i$ or $\neg\varphi \in \Phi_i$; 2) *consistency*: there is no φ such that $\varphi \in \Phi_i$ and $\neg\varphi \in \Phi_i$; 3) *deductive closure* with respect to the agenda: $\mathrm{Cn}(\Phi_i) \cap \Phi \subseteq \Phi_i$.

A *collective judgment* is a subset $\Psi \subseteq \Phi$ such that Ψ also satisfies the rationality constraints and (in some sense) it is *a response* to the individual judgments of all and only the agents from $\mathcal{A}$. The desired properties of responsiveness are characterized by three requirements imposed on the judgment aggregation function (JAF), i.e. a function that, given a profile of all individual judgments $\{\Phi_i\}_{i\in\mathcal{A}}$, should uniquely determine the collective judgment. These are: 1) *universal domain*: a JAF should yield a collective judgment for any possible profile of individual judgments; 2) *anonymity*: the individual judgments of agents should have equal importance in determining the outcome; 3) *independence*:[2] for every proposition φ and any two profiles of individual judgments, if for every $i \in \mathcal{A}$ it holds that $\varphi \in \Phi_i$ iff $\varphi \in \Phi'_i$ then $\varphi \in \mathrm{JAF}(\{\Phi_i\}_{i\in\mathcal{A}})$ iff $\varphi \in \mathrm{JAF}(\{\Phi'_i\}_{i\in\mathcal{A}})$.

A seemingly natural choice for a JAF is the *propositionwise majority voting rule*, which advises accepting collectively all and only those propositions from Φ that are accepted individually by the majority of agents. As it turns out, however, if Φ contains at least two different propositions and their conjunction, the majority procedure may lead to obtaining inconsistent collective judgments. This effect is known as the *discursive dilemma* or *doctrinal paradox* and has recently attracted much attention in the field of computational social choice, e.g. [12, 1, 2, 14]. The dilemma, as List and Pettit [12] have proved, is unavoidable under the six previously listed constraints. Nevertheless, there is a number of ways of escaping it by relaxing some of the requirements. In the same paper the authors present a comprehensive discussion of different solutions, of which we shall mention two.

The first one avoids the paradox by dropping the completeness requirement on the collective judgment. Under specific provisions a group might simply suspended its judgment on particular propositions. This way the outcome is deductively weaker and does not provide a full solution to the given aggregation task, but inconsistency does not occur. The other method resolves the paradox by relaxing the independence assumption through conditioning acceptance of certain propositions on the judgment on some others. Two typically invoked strategies include the premise- and the conclusion-driven aggregation, according to which the priority is given to those propositions that serve respectively as the premises or the conclusion in the agenda. The judgment on the remaining ones is then suitably adjusted in order to avoid inconsistency.

The basic shortcoming of both approaches is their inability of resolving the paradox in an unambiguous and nonarbitrary manner. For instance, for the same input the premise- and the conclusion-driven procedures can easily yield incompatible outcomes, while none of them is more intuitive than the other.

Recently, an argument-driven approach, inspired by an operation of merging belief bases in AI [7, 8], has been also investigated as a strategy of relaxing independence [14, 15]. The

[2] The *independence* condition is weaker than the original *systematicity* requirement used in [12] and so, as slightly less controversial, tends to replace the former constraint in more recent literature, e.g. [1].

method is considerably better justified and well-behaved than the aforementioned procedures. It employs a simple distance measure of individual judgments from possible consistent collective judgments, and chooses the one (or those) that minimize it. Thus the preference is given not to the premises or the conclusion but to the argument as a whole. Our proposal rests upon similar principles, but instead builds on certain results from philosophy of science and significantly generalizes the approach.

The problem of judgment aggregation, interpreted as a specific case of propositional acceptance, can be related to a similar question of how to aggregate logically connected hypotheses about the world into a coherent body of scientific knowledge. Interestingly, a direct analogue of the discursive dilemma occurs also in that context [11, 2] and has been a recurrent subject of debates and analyses in the XXth century philosophy of science, e.g. [5, 9]. Essentially, the *lottery paradox* shows that propositionwise acceptance over logically connected statements fails in general in yielding a consistent set of formulas. According to I. Levi [9], the problem stems from a too narrow perspective on acceptance in science. Scientific inquiry is a goal-oriented activity, driven by (at least) two rival goals of a purely epistemic nature: obtaining *true* and highly *informative* statements. Any plausible conclusion of an inference may satisfy them to different extent, and so be relatively better or worse with respect to other candidate answers. If the scientist is able to assess the degree to which the goals are met by particular conclusions and possesses some probabilistic knowledge about the possible states of the world, then he should evaluate the *expected epistemic utility* of the conclusions and simply accept the one that *maximizes* it. The *cognitive decision model*, proposed by Levi, aims to give a formal account of such a mechanism of acceptance.

The aggregation method presented in the following section is an extension of Levi's model. It borrows its all basic assumptions and the chief part of its formal apparatus. The novel share involves defining a measure of the social agreement, a method of generating probability distributions from profiles of individual judgments, and restating the judgment aggregation problem as a task of satisfying (often rival) epistemic and social goals. Some limited experimental results are presented in Section 4.

3 The Utilitarian Model of Judgment Aggregation

A group of agents striving to construct a collective judgment on some issue wants the judgment to have certain good properties. Namely, it has to reflect individual judgments of the group's members and moreover it has to be a rational statement by itself. The former requirement, to which discussions on judgment aggregation have been mainly confined, involves applying some measure of responsiveness of the collective judgment to individual beliefs of agents. The latter rests upon the assumption that a collective judgment quite often conveys a particular claim about the world, which can be evaluated with respect to the epistemic objectives mentioned above.

To start with, the utilitarian model of aggregation requires recognition of the set of all possible states of the world associated, for instance, with the logical models of the agenda. Consider $\Phi = \{p, \neg p, q, \neg q, r, \neg r\}$ and the background knowledge[3] $b = \{p \wedge q \leftrightarrow r\}$. Under the given constraints there are four distinct truth valuation functions over the formulas from Φ, corresponding to four complete, consistent and deductively closed sets of judgments on Φ: $\{p, q, r\}, \{\neg p, q, \neg r\}, \{p, \neg q, \neg r\}, \{\neg p, \neg q, \neg r\}$. We shall interpret the collection of these functions $\mathcal{M}_{\Phi,b} = \{v_1, ..., v_4\}$ as the set of all possible and mutually exclusive ways the world might be with respect to Φ and b. By convention we posit that judgment Ψ is satisfied by state v_i, i.e. $v_i \vDash \Psi$, whenever v_i makes all formulas in Ψ true.

[3]The assumption of background knowledge is used only to abbreviate the notation, and can be dropped at any time by replacing atomic formulas in Φ by respective compound ones, defined as in b.

Further, it is necessary to specify the *answer set*, i.e. the set of all collective judgments whose acceptance could present certain value to the voting group. Typically, the judgments satisfying the rationality constraints should be permitted in the first order as possible outcomes of the aggregation procedure. Also, in many cases, a group might want to relax some of the requirements — predominantly completeness — to extend the range of possible outcomes. The following partial, though still consistent and deductively closed judgments on Φ could be often seen as interesting in a variety of contexts: $\{\{p\}, \{q\}, \{\neg p, \neg r\}, \{\neg q, \neg r\}, \{\neg r\}\}$. For instance, if r is a legal verdict based upon two premises p and q, it could be enough for the jury to agree on the negation of only one of the premises, since this alone allows for determining the conclusion $\neg r$. Even if the jury cannot decide on the truth value of the second premise, it can still make a final judgment and justify it.

Partial judgments can be evaluated with respect to the amount of information they convey. By analogy to the cognitive decision model, this can be defined in terms of the proportion of possible states of the world that are excluded by the given judgment:[4]

$$\mathrm{cont}(\Psi) = \frac{|v_i \in \mathcal{M}_{\Phi,b} : v_i \not\models \Psi|}{|\mathcal{M}_{\Phi,b}|}$$

In the example under discussion we obtain:

$$\mathrm{cont}(\{p, q, r\}) = 0.75 \qquad \mathrm{cont}(\{p\}) = 0.5 \qquad \mathrm{cont}(\{\neg r\}) = 0.25$$

Measure cont, based on purely structural properties of the problem, can be suitably parameterized to account also for additional pragmatic value that a judgment offers to the group. Figure 1 shows three sample ways of evaluating information relative to the proportion of possible states that are excluded by the judgment. Notice, that whereas the linear plot represents the standard measure, the two others may be adopted by a group revealing a weaker (top) or a stronger (bottom) bias for completeness of the outcome.

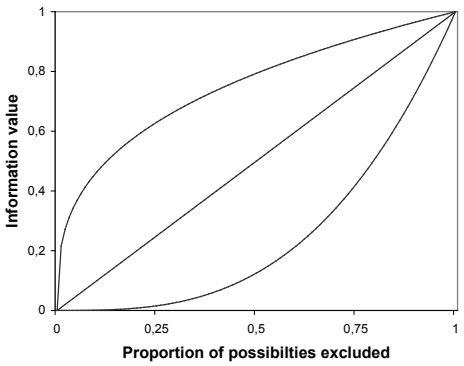


Figure 1: The value of information with respect to information content.

Another possibility of extending the range of permissible judgments regards dropping the consistency requirement. According to D. Fallis [3] inconsistent judgments do not have to be always worthless. The information content of an inconsistent judgment should by definition of cont be equal 1, which neatly harmonizes with the property of inferential explosion. However, a group might want to assign other values, according to its particular understanding of informativeness of an inconsistent statement. Also judgments that are not deductively closed can be formally incorporated into the model, if only the group can assess

[4]Other interesting measures of information content are discussed in e.g. D. Fallis [4] and P. Maher [13].

their information value in a meaningful sense. Finally, $\emptyset$ with $\mathrm{cont}(\emptyset) = 0$ is worth including in the answer set as an option for suspending the judgment, which often might be the only reasonable decision.

Let then $\mathcal{CJ} = \{\Psi_1, \ldots, \Psi_m\}$ be a set of possible collective judgments, preselected and evaluated with respect to information content by the group. Following the model of cognitive decision we will assess the value of truth — the second epistemic objective traded off against the information content — relatively to the state of the world, using a simple binary measure:

$$\mathrm{T}(\Psi, v_i) = \begin{cases} 1 \text{ iff } v_i \vDash \Psi \\ 0 \text{ iff } v_i \nvDash \Psi \end{cases}$$

The measure takes value 1 in state v_i, whenever a judgment is true in it, and 0 otherwise. The overall utility of accepting a collective judgment, provided that state v_i holds, is given by the following function, which assigns numerical values to potential judgments according to the degree to which they satisfy the two epistemic goals:

$$u_\varepsilon(\Psi, v_i) = \alpha \ \mathrm{cont}(\Psi) + (1 - \alpha) \ \mathrm{T}(\Psi, v_i)$$

Coefficient $\alpha \in [0, 1]$ serves here as an epistemic preference indicator, with 0 standing for the full preference for truth, and 1 for the information content of a judgment.

In order to complete the framework one has to employ knowledge about the probability of the possible states. Although in a typical judgment aggregation problem no such information is explicitly provided, there is a way of inducing a probability distribution using the profile of individual judgments. Notice, that every individual judgment, as satisfying the rationality constraints, corresponds to exactly one model from $\mathcal{M}_{\Phi,b}$. Assuming that agents are characterized by a certain degree of reliability, it is justified to regard their judgments as good indicators of the truth of the states (cf. [15]). Formally, this information can determine the probability distribution over $\mathcal{M}_{\Phi,b}$ by means of Bayes' Theorem.

First, a uniform prior distribution is posed over the states, i.e. $P(v_i) = \frac{1}{|\mathcal{M}_{\Phi,b}|}$ for every $v_i \in \mathcal{M}_{\Phi,b}$. Let r, such that $0.5 > r > 1$, be the degree of reliability of agents, meaning that given state v_i is true, the probability $P(\Phi|v_i)$ that an agent makes a correct judgment ($v_i \vDash \Phi$) is r, whereas the probability of the opposite is $1 - r$. Starting with the prior probabilities every individual judgment is used to update the distribution over $\mathcal{M}_{\Phi,b}$, so that posterior probability of a state, given an individual judgment Φ, equals to:

$$P(v_i|\Phi) = \frac{P(\Phi|v_i)P(v_i)}{\sum_j P(\Phi|v_j)P(v_j)}$$

It can be shown that the resulting distribution after n consecutive updates, i.e. for n individual judgments, is given by the equation:

$$P^\star(v_i) = \frac{r^{n_i}(1-r)^{n-n_i}}{\sum_i r^{n_i}(1-r)^{n-n_i}}$$

where $\sum_i n_i = n$ and each n_i is the number of supporters of state v_i. A distribution generated in this way has two interesting properties. First, it reflects the beliefs of agents to the degree that these beliefs can be deemed correct, thus complying to the principles of the Bayesian epistemology. Judgments of unreliable agents (for r approaching 0.5) do not have a strong influence on the distribution, as opposed to those of highly reliable judges (for r being close to 1), which are given maximal weight. Second, for a finite number of agents none of the possible states is ever completely excluded, i.e. $P^\star(v_i) > 0$ for all $v_i \in \mathcal{M}_{\Phi,b}$, hence the fundamental uncertainty, which often motivates the need for judgment aggregation, is not eliminated totally.

Up to this point the model roughly mimics the design of Levi's framework, giving account of the epistemic aspects of acceptance that, as we argue, play essential role in the problem

of aggregating judgments. The central claim of this paper, however, is that an aggregation procedure is driven by the endeavor to maximize both epistemic and social benefits following from acceptance of a collective judgment. Epistemically good judgments are not fully satisfactory unless they also justly reflect opinions of the agents involved. Intuitively, socially the fairest collective judgment is the one selected by propositionwise majoritarian rule. Nevertheless, as already pointed out, this strategy cannot be reliably employed. Instead, we should allow the choice to be made only from the answer set $\mathcal{CJ}$. For this purpose we will use a measure of social agreement interpreted as a form of distance of a judgment from the majoritarian choice. Let SA be a function whose values are assigned as follows:

- for any $\varphi \in \Phi$: $\mathrm{SA}(\varphi) = \frac{|\mathcal{A}_\varphi|}{|\mathcal{A}|}$, i.e. the social agreement on a proposition is the ratio of the number of agents who accept it to the total number of agents;
- for any $\Psi \in \mathcal{CJ}$: $\mathrm{SA}(\Psi) = \frac{1}{|\Psi|} \sum_{\varphi \in \Psi} \mathrm{SA}(\varphi)$, i.e. the social agreement on a collective judgment is equal to the arithmetic mean of the degrees of social agreement on propositions accepted in the judgment.

SA expresses what proportion of propositions from a judgment is on average accepted by an agent. We thus stay very close to the notion of Hamming distance proposed as the basis for the argument-driven approach to judgment aggregation [14].[5] The suspension of judgment can be assigned value 0, as it is in no way responsive to individual judgments.

SA, as considered independently from the rest of the model, shares interesting similarities with the aforementioned aggregation procedures that rest on the relaxation of the independence constraint. If we restrict $\mathcal{CJ}$ to contain all and only complete sets of judgments, including the inconsistent ones, then the judgment Ψ such that $\mathrm{SA}(\Psi) = \max_i \mathrm{SA}(\Psi_i)$ is equivalent to the choice of the propositionwise majority voting rule. This is due to the fact, that out of all complete judgments, the one with maximum value of SA is the one whose every accepted proposition has $\mathrm{SA}(\varphi) \geq 0.5$. If inconsistent judgments are left out, then the one that maximizes the degree of social agreement can be seen as socially the closest in $\mathcal{CJ}$ to the majoritarian choice. This resolves the dilemma of arbitrary selection between the premise- and the conclusion-driven strategy. The measure SA alone points at the judgment that violates the majoritarian vote to the smallest degree, no matter whether violation occurs on the side of premises or conclusions. Consider again the example with $\Phi = \{p, \neg p, q, \neg q, r, \neg r\}$ and $b = \{p \wedge q \leftrightarrow r\}$. If the profile of individual judgments is such that the following degrees of social agreement are assigned to the propositions:

$$\mathrm{SA}(p) = 0.8 \qquad \mathrm{SA}(q) = 0.7 \qquad \mathrm{SA}(r) = 0.4$$

then the socially best, complete and consistent judgment is $\{p, q, r\}$, since the value of r is the closest to 0.5, and so accepting r, even though it is against the majoritarian vote, brings the least disagreement amongst agents. The choice is therefore equivalent to the outcome of the premise-driven majoritarian procedure. In case of another distribution of agreement, for instance:

$$\mathrm{SA}(p) = 0.7 \qquad \mathrm{SA}(q) = 0.6 \qquad \mathrm{SA}(r) = 0.3$$

the judgment $\{p, \neg q, \neg r\}$ would be the most preferable, which is compatible with the conclusion-driven strategy.

Finally, the epistemic and social factors can be embraced within a single utility function defined for any $\Psi \in \mathcal{CJ}$ and $v_i \in \mathcal{M}_\Phi$ as:

$$\begin{aligned} u(\Psi, v_i) &= \beta \underbrace{(\alpha\ \mathrm{cont}(\Psi) + (1-\alpha)\ \mathrm{T}(\Psi, v_i))}_{u_\varepsilon(\Psi, v_i)} + (1-\beta)\ \mathrm{SA}(\Psi) \\ &= \beta \qquad\qquad\qquad\qquad\quad + (1-\beta)\ \mathrm{SA}(\Psi) \end{aligned}$$

[5] In fact SA is a normalized form of the measure adopted by G. Pigozzi in [14].

While α controls the trade-off on the purely epistemic level, coefficient $\beta \in [0,1]$ is supposed to reflect the upper-level preferences of a group, balancing the trade-off between the epistemic and social perspective. For β approaching 1, the epistemic criteria take over and an act of judgment aggregation becomes an act of rational acceptance, analogous to the typical cases modeled in Levi's framework. When it is close to 0, the social agreement measure becomes a decisive factor, rendering the procedure majoritarian, but still free of paradoxes. If the group is able to correctly diagnose its preferences and set the values of α and β accordingly, then the judgment maximizing the expected utility, i.e.:

$$\mathrm{EU}(\Psi) = \textstyle\sum_{v_i \in \mathcal{M}_\Phi} P^\star(v_i) u(\Psi, v_i)$$

should be the one to be rationally accepted, as satisfying to the greatest extent the collective preferences of the group. For a complete picture, the procedure requires a tie-breaking rule. The prescription proposed by Levi [9], namely to accept the disjunction of all the answers with maximum expected utility, could be interpreted in this context as acceptance of the *common information* included in the judgments maximizing the expected utility. If we provisionally accept such a rule,[6] we can conclude the presentation with the utilitarian judgment aggregation function, formulated as follows:

$$\text{(UJAM)} \qquad \mathrm{JAF}(\{\Phi_i\}_{i\in\mathcal{A}}) = \bigcap \Psi \quad \text{such that} \quad \Psi \in \arg\max_{\Psi \in \mathcal{CJ}} \mathrm{EU}(\Psi)$$

$\star$

Before concluding the presentation we will summarize the model's ingredients introduced in this section. Given a judgment aggregation problem specified by a set of agents $\mathcal{A}$, a set of propositional formulas Φ and a profile of individual judgments $\{\Phi_i\}_{i\in\mathcal{A}}$, where each judgment has to be a complete, consistent, and deductively closed subset of Φ, the group can define utilitarian aggregation model consisting of the following elements:

$\mathcal{CJ} = \{\Psi_1, ..., \Psi_m\} \subseteq 2^\Phi$	the answer set: a set of potential collective judgments preselected by the group as presenting certain value,
$\mathcal{M}_\Phi = \{v_1, ..., v_l\}$	the set of all possible, mutually exclusive states of the world (models) with respect to Φ,
$v_i \in \mathcal{M}_\Phi : \Phi \to \{0,1\}$	a truth valuation function on the propositions from Φ,
$P^\star : \mathcal{M}_\Phi \to [0,1]$	a probability function over possible states, whose values are derived from the profile of individual judgments,
$r \in (0.5, 1)$	the degree of reliability of individual judgments / agents,
$\alpha, \beta \in [0,1]$	the coefficients controlling the information-truth and epistemic-social trade-offs,
$\mathrm{cont} : \mathcal{CJ} \to [0,1]$	the valuation measure of information contained in the collective judgments,
$\mathrm{T} : \mathcal{CJ} \times \mathcal{M}_\Phi \to \{0,1\}$	the valuation measure of truth with respect to the collective judgments and possible states of the world,
$\mathrm{SA} : \mathcal{CJ} \cup \Phi \to [0,1]$	the measure of social agreement on the propositions and collective judgments.

[6] Clearly, the outcome of such a rule does not have to necessarily maximize the expected utility. Note also, that alternative interpretations of Levi's proposal, though less appealing, are possible. Yet other options of establishing a tie-breaking rule include typical decision-theoretic escape routes, for instance: 1) providing an a priori preference ranking over the goals, so that the judgment which satisfies the most preferred goal better is picked; 2) employing another decision criterion (e.g. Hurwicz, Laplace), which have chances to yield a unique outcome. Both solutions suffer, however, from the same arbitrariness as was previously pointed out in the standard aggregation methods.

4 Experimentation

In this section we will present sample aggregation results obtained with the utilitarian judgment aggregation model for set of propositions $\Phi = \{p, \neg p, q, \neg q, r, \neg r\}$ and background knowledge $b = \{p \wedge q \leftrightarrow r\}$. In the scenario there are four possible states of the world (models): $\mathcal{M}_{\Phi,b} = \{v_1, v_2, v_3, v_4\}$, defined by the following truth valuations:

$$\begin{array}{ll} v_1: & v_1(p) = 1,\ v_1(q) = 1,\ v_1(r) = 1 \\ v_2: & v_2(p) = 0,\ v_2(q) = 1,\ v_2(r) = 0 \\ v_3: & v_3(p) = 1,\ v_3(q) = 0,\ v_3(r) = 0 \\ v_4: & v_4(p) = 0,\ v_4(q) = 0,\ v_4(r) = 0 \end{array}$$

The value of information content for all collective judgments Ψ is assigned according to the standard measure $\text{cont}(\Psi)$. The shaded rows in the tables below mark the collective judgments selected by the model.

Information content vs. truth: Tables 1 - 2. For a constant profile of individual judgments and a fixed value of the epistemic-social coefficient $\beta = 0.8$, we change the value of the information-truth coefficient α. For $\alpha = 0.7$ (Table 1) $\{p, \neg q, \neg r\}$ is chosen as the most informative judgment among those which are still sufficiently probable and agreeable. Decreasing the value of α results gradually in selection of less informative (more incomplete) collective judgments. For $\alpha = 0.5$ we get $\{\neg q, \neg r\}$, whereas for $\alpha = 0.1$, $\{\neg r\}$ (Table 2). When fixed to $\alpha = 0.01$ the aggregation becomes so truth-oriented that only the suspension of judgment is a plausible choice.

Information content vs. social agreement: Tables 3 - 4. For a constant profile of individual judgments we fix the value of the information-truth coefficient at $\alpha = 1$, meaning that truth is completely excluded from considerations. Shifting the value of the epistemic-social coefficient from $\beta = 0.5$ (Table 3) to $\beta = 0.3$ (Table 4), leads to the change in the accepted collective judgment from $\{p, \neg q, \neg r\}$ to more agreeable though incomplete $\{p\}$.

Truth vs. social agreement: Tables 5 - 6. For a constant profile of individual judgments we fix the value of the information-truth coefficient at $\alpha = 0$, thus discarding information content. Setting $\beta = 0.6$ (Table 5) we reveal a higher preference for a conclusion more likely to be true, which in this case is $\{\neg r\}$. Notice, that $\{\neg r\}$ has the same degree of agreement as $\{\neg q\}$ and even lower than $\{p\}$, but $\{\neg r\}$ is true in three out of four possible states with

Total Number of Agents:	10	Individual Judgments:	2	1	4	3	
Reliability of Agents:	0.55	Probability:	0.22	0.18	0.33	0.27	
Collective Judgments:	Inform. Value:	Degree of Agreement:	v_1	v_2	v_3	v_4	Expected Utility:
$\{p,\ q,\ r\}$	0.75	0.367	0.733	0.493	0.493	0.493	0.546
$\{\neg p,\ q, \neg r\}$	0.75	0.500	0.520	0.760	0.520	0.520	0.563
$\{p, \neg q, \neg r\}$	0.75	0.700	0.560	0.560	0.800	0.560	0.639
$\{\neg p, \neg q, \neg r\}$	0.75	0.633	0.547	0.547	0.547	0.787	0.611
$\{\neg p, \neg r\}$	0.50	0.600	0.400	0.640	0.400	0.640	0.508
$\{\neg q, \neg r\}$	0.50	0.750	0.430	0.430	0.670	0.670	0.574
$\{p\}$	0.50	0.600	0.640	0.400	0.640	0.400	0.532
$\{q\}$	0.50	0.300	0.580	0.580	0.340	0.340	0.436
$\{\neg r\}$	0.25	0.800	0.300	0.540	0.540	0.540	0.487
suspend	0	0	0.240	0.240	0.240	0.240	0.240

Degrees of Agreement on Atoms:		
	p	0.6
	q	0.3
	r	0.2

Trade-offs:	
Information vs. Truth	0.7
Epistemic vs. Social	0.8

Table 1.

Total Number of Agents:	10	Individual Judgments:	2	1	4	3	
Reliability of Agents:	0.55	Probability:	0.22	0.18	0.33	0.27	
Collective Judgments:	Inform. Value:	Degree of Agreement:	v_1	v_2	v_3	v_4	Expected Utility:
$\{p,\ q,\ r\}$	0.75	0.367	0.853	0.133	0.133	0.133	0.292
$\{\neg p,\ q, \neg r\}$	0.75	0.500	0.160	0.880	0.160	0.160	0.290
$\{p, \neg q, \neg r\}$	0.75	0.700	0.200	0.200	0.920	0.200	0.437
$\{\neg p, \neg q, \neg r\}$	0.75	0.633	0.187	0.187	0.187	0.907	0.381
$\{\neg p, \neg r\}$	0.50	0.600	0.160	0.880	0.160	0.880	0.484
$\{\neg q, \neg r\}$	0.50	0.750	0.190	0.190	0.910	0.910	0.621
$\{p\}$	0.50	0.600	0.880	0.160	0.880	0.160	0.556
$\{q\}$	0.50	0.300	0.820	0.820	0.100	0.100	0.389
$\{\neg r\}$	0.25	0.800	0.180	0.900	0.900	0.900	0.741
suspend	0	0	0.720	0.720	0.720	0.720	0.720

Degrees of Agreement on Atoms:		
	p	0.6
	q	0.3
	r	0.2

Trade-offs:	
Information vs. Truth	0.1
Epistemic vs. Social	0.8

Table 2.

Total Number of Agents:	10	Individual Judgments:	4	1	4	1	
Reliability of Agents:	0.6	Probability:	0.39	0.11	0.39	0.11	
Collective Judgments:	Inform. Value:	Degree of Agreement:	v_1	v_2	v_3	v_4	Expected Utility:
$\{p,\ q,\ r\}$	0.75	0.567	0.658	0.658	0.658	0.658	0.658
$\{\neg p,\ q, \neg r\}$	0.75	0.433	0.592	0.592	0.592	0.592	0.592
$\{p, \neg q, \neg r\}$	0.75	0.633	0.692	0.692	0.692	0.692	0.692
$\{\neg p, \neg q, \neg r\}$	0.75	0.433	0.592	0.592	0.592	0.592	0.592
$\{\neg p, \neg r\}$	0.50	0.400	0.450	0.450	0.450	0.450	0.450
$\{\neg q, \neg r\}$	0.50	0.550	0.525	0.525	0.525	0.525	0.525
$\{p\}$	0.50	0.800	0.650	0.650	0.650	0.650	0.650
$\{q\}$	0.50	0.500	0.500	0.500	0.500	0.500	0.500
$\{\neg r\}$	0.25	0.600	0.425	0.425	0.425	0.425	0.425
suspend	0	0	0.000	0.000	0.000	0.000	0.000

Degrees of Agreement on Atoms:		
	p	0.8
	q	0.5
	r	0.4

Trade-offs:	
Information vs. Truth	1
Epistemic vs. Social	0.5

Table 3.

Total Number of Agents:	10	Individual Judgments:	4	1	4	1	
Reliability of Agents:	0.6	Probability:	0.39	0.11	0.39	0.11	
Collective Judgments:	Inform. Value:	Degree of Agreement:	v_1	v_2	v_3	v_4	Expected Utility:
$\{p,\ q,\ r\}$	0.75	0.567	0.622	0.622	0.622	0.622	0.622
$\{\neg p,\ q, \neg r\}$	0.75	0.433	0.528	0.528	0.528	0.528	0.528
$\{p, \neg q, \neg r\}$	0.75	0.633	0.668	0.668	0.668	0.668	0.668
$\{\neg p, \neg q, \neg r\}$	0.75	0.433	0.528	0.528	0.528	0.528	0.528
$\{\neg p, \neg r\}$	0.50	0.400	0.430	0.430	0.430	0.430	0.430
$\{\neg q, \neg r\}$	0.50	0.550	0.535	0.535	0.535	0.535	0.535
$\{p\}$	0.50	0.800	0.710	0.710	0.710	0.710	0.710
$\{q\}$	0.50	0.500	0.500	0.500	0.500	0.500	0.500
$\{\neg r\}$	0.25	0.600	0.495	0.495	0.495	0.495	0.495
suspend	0	0	0.000	0.000	0.000	0.000	0.000

Degrees of Agreement on Atoms:		
	p	0.8
	q	0.5
	r	0.4

Trade-offs:	
Information vs. Truth	1
Epistemic vs. Social	0.3

Table 4.

almost uniform probability, and so it is more probable than the other two. When we fix $\beta = 0.4$ (Table 6) the accepted judgment is $\{p\}$ as more agreeable. The same outcome would be yielded for $\beta = 0.6$ were the agents more reliable, say for $r = 0.7$. In that case the probability distribution would be strongly influenced by individual judgments and the two states where $\{p\}$ is true would become much more probable than the others.

Total Number of Agents:	10	Individual Judgments:	4	0	4	2	
Reliability of Agents:	0.51	Probability:	0.26	0.23	0.26	0.24	
Collective Judgments:	Inform. Value:	Degree of Agreement:	v_1	v_2	v_3	v_4	Expected Utility:
$\{p,\ q,\ r\}$	0.75	0.533	0.813	0.213	0.213	0.213	0.372
$\{\neg p,\ q, \neg r\}$	0.75	0.400	0.160	0.760	0.160	0.160	0.295
$\{p, \neg q, \neg r\}$	0.75	0.667	0.267	0.267	0.867	0.267	0.426
$\{\neg p, \neg q, \neg r\}$	0.75	0.467	0.187	0.187	0.187	0.787	0.333
$\{\neg p, \neg r\}$	0.50	0.400	0.160	0.760	0.160	0.760	0.442
$\{\neg q, \neg r\}$	0.50	0.600	0.240	0.240	0.840	0.840	0.546
$\{p\}$	0.50	0.800	0.920	0.320	0.920	0.320	0.638
$\{q\}$	0.50	0.400	0.760	0.760	0.160	0.160	0.454
$\{\neg r\}$	0.25	0.600	0.240	0.840	0.840	0.840	0.681
suspend	0	0	0.600	0.600	0.600	0.600	0.600

Degrees of Agreement on Atoms:		
	p	0.8
	q	0.4
	r	0.4

Trade-offs:	
Information vs. Truth	0
Epistemic vs. Social	0.6

Table 5.

Total Number of Agents:	10	Individual Judgments:	4	0	4	2	
Reliability of Agents:	0.51	Probability:	0.26	0.23	0.26	0.24	
Collective Judgments:	Inform. Value:	Degree of Agreement:	v_1	v_2	v_3	v_4	Expected Utility:
$\{p,\ q,\ r\}$	0.75	0.533	0.720	0.320	0.320	0.320	0.426
$\{\neg p,\ q, \neg r\}$	0.75	0.400	0.240	0.640	0.240	0.240	0.330
$\{p, \neg q, \neg r\}$	0.75	0.667	0.400	0.400	0.800	0.400	0.506
$\{\neg p, \neg q, \neg r\}$	0.75	0.467	0.280	0.280	0.280	0.680	0.378
$\{\neg p, \neg r\}$	0.50	0.400	0.240	0.640	0.240	0.640	0.428
$\{\neg q, \neg r\}$	0.50	0.600	0.360	0.360	0.760	0.760	0.564
$\{p\}$	0.50	0.800	0.880	0.480	0.880	0.480	0.692
$\{q\}$	0.50	0.400	0.640	0.640	0.240	0.240	0.436
$\{\neg r\}$	0.25	0.600	0.360	0.760	0.760	0.760	0.654
suspend	0	0	0.400	0.400	0.400	0.400	0.400

Degrees of Agreement on Atoms:		
	p	0.8
	q	0.4
	r	0.4

Trade-offs:	
Information vs. Truth	0
Epistemic vs. Social	0.4

Table 6.

5 Conclusions and Discussion

The judgment aggregation model introduced in this paper aims at capturing the concept of the utility that an act of acceptance of a collective judgment can offer to a group of rational agents. For this purpose we have reinterpreted the problem of aggregation as a goal-oriented task with (possibly rival) goals of an epistemic and social character involved. Following Levi's cognitive decision model [9, 10], we have identified the epistemic goals as truth and

information content and adopted the respective measures of the degrees to which these goals are satisfied by potential collective judgments. Further, we have defined a measure of the agreement on a judgment, which reflects the social objective of the procedure. Finally, an utility function defined as a weighted sum of the three measures has been proposed, together with an acceptance rule based on the criterion of maximizing the expected utility.

As a method of aggregation the model imposes two requirements on a group: designating the answer set and expressing numerically its preferences with respect to the goals involved. Assuming that this is done correctly, the collective judgment that is selected by the model is guaranteed to be the best one among all candidate judgments.

Clearly, the model might not satisfy the six requirements normally imposed on judgment aggregation function and the resulting collective judgments. However, any violation of these receives here a strong justification.

Completeness, consistency, deductive closure: As long as only complete, consistent and deductively closed collective judgments are considered, none of the requirements will be violated by the outcome (provided that tie does not occur). If, on the contrary, the answer set contains other judgments as well, then there is no good reason to defend the constraints. The doctrinal paradox obtains, therefore, a straightforward solution. If the group finds an inconsistent collective judgment undesirable, it should not designate it as a potential outcome; otherwise its occurrence is apparently not troublesome.

Universal domain, anonymity: For any profile of judgments the model designates a unique outcome (assuming the tie-breaking rule is employed) and assigns the same weight to every individual judgment in determining it.

Independence: This constraint can often be not satisfied by the model. Still, the central rationale behind the presented framework is a conviction that independence is a too strong requirement. Investigations into the discursive dilemma and the lottery paradox show that propositionwise acceptance rules — the type of rules enforced by the independence constraint — inevitably lead to inconsistencies when applied to sets of logically connected propositions. Following Levi, we argue that only statements considered in the context of the entire logical structure to which they belong can be rationally accepted or rejected. As a consequence a proposition obtaining the support of exactly the same agents can be once accepted, while rejected another time. This, however, happens only because the judgments containing this proposition are evaluated differently in the two cases, and therefore, in a broader perspective it is clearly a desired effect.

As a theoretical framework the model is universal enough to be amenable to many interesting adjustments and extensions, thus providing a large space of possible applications. Its most important advantage is that it offers a good control over the trade-offs involved in the aggregation task. Predominantly, it gives a clear formal account of how the responsiveness of the procedure — the fundamental social choice postulate — can be weighted against other expectations regarding collective judgments.

From the practical perspective a serious deficiency of the method, which should be a subject to further analysis, concerns its computational tractability. Depending on the structure of the problem it might be necessary to consider up to $2^{|\Phi|}$ possible states and the same number of candidate collective judgments. In any case, significant savings on the computational expense can be achieved at the cost of dropping one or two measures involved. Two scenarios seem especially appealing in this respect: 1) *dropping truth* (and consequently probabilities), which dramatically reduces the computational effort, while still allows for tracing the trade-off between social agreement and completeness of collective judgments; 2) *dropping epistemic utility*, which turns the method into a robust majoritarian procedure, selecting the permissible judgment that is closest to the majoritarian choice (an approach investigated by G. Pigozzi in [14]).

Acknowledgments. I would like to thank Ulle Endriss for his helpful feedback on the initial version of this paper, which has been written while studying the *Master of Logic* programme at the Institute for Logic, Language and Computation, University of Amsterdam. I am also grateful to three anonymous reviewers at COMSOC-2008 for valuable comments and suggestions.

References

[1] F. Dietrich, A generalised model of judgment aggregation. In *Social Choice and Welfare* 28, pp. 529-565, 2007.

[2] I. Douven, JW. Romeijn, The Discursive Dilemma as a Lottery Paradox. In U. Endriss, J. Lang, *Proceedings of the 1st International Workshop on Computational Social Choice (COMSOC-2006)*, ILLC University of Amsterdam, pp. 164-177, 2006.

[3] D. Fallis, Epistemic Value Theory and Judgment Aggregation. In *Episteme* 2(1), pp. 39-55, 2005.

[4] D. Fallis, Measures of Epistemic Utility and the Value of Experiments. In *Proceedings of the 17th Biennial Meeting of the Philosophy of Science Association*, Vancouver 2000.

[5] C. G. Hempel, Inductive Inconsistencies. In *Synthese* 12(4), pp. 439-469, 1960.

[6] R. Jeffrey, Valuation and Acceptance of Scientific Hypotheses. In *Philosophy of Science* 23(3), pp. 237-246, 1956.

[7] S. Konieczny, R. P. Pérez, Merging information under constraints: a logical framework, *Journal of Logic and Computation* 12(5), pp. 773-808, 2002.

[8] S. Konieczny, R. P. Pérez, Propositional belief base merging or how to merge beliefs/goals coming from several sources and some links with Social choice theory. In *European Journal of Operational Research* 160(3), pp. 785-802, 2005.

[9] I. Levi, *Gambling with Truth. An Essay on Induction and the Aims of Science*, MIT Press: Cambridge, 1967.

[10] I. Levi, Information and Inference. In *Synthese* 17(1), pp. 369-391, 1967.

[11] I. Levi, List and Pettit. In *Synthese* 140, pp. 237-242, 2004.

[12] C. List, P. Pettit, Aggregating Sets of Judgments: an Impossibility Result. In *Economics and Philosophy* 18, pp. 89-110, 2002.

[13] P. Maher, *Betting on Theories*, Cambridge University Press: Cambridge, 1996.

[14] G. Pigozzi, Belief merging and the discursive dilemma: an argument-based account to paradoxes of judgment aggregation. In *Synthese* 152(2), pp. 285-298, 2006.

[15] G. Pigozzi, S. Hartmann, Merging Judgments and the Problem of Truth-Tracking. In U. Endriss, J. Lang, *Proceedings of the 1st International Workshop on Computational Social Choice (COMSOC-2006)*, ILLC University of Amsterdam, pp. 408-421, 2006.

Szymon Klarman
ILLC, University of Amsterdam
1018 TV Amsterdam, The Netherlands
Email: `sklarman@science.uva.nl`

Confluence Operators: Negotiation as Pointwise Merging [1]

Sébastien Konieczny
CRIL - CNRS
Université d'Artois, Lens, France
konieczny@cril.fr

Ramón Pino Pérez
Departamento de Matemáticas, Facultad de ciencias
Universidad de Los Andes, Mérida, Venezuela
pino@ula.ve

Abstract

In the logic based framework of knowledge representation and reasoning many operators have been defined in order to capture different kinds of change: revision, update, merging and many others. There are close links between revision, update, and merging. Merging operators can be considered as extensions of revision operators to multiple belief bases. And update operators can be considered as pointwise revision, looking at each model of the base, instead of taking the base as a whole. Thus, a natural question is the following one: Are there natural operators that are pointwise merging, just as update are pointwise revision? The goal of this work is to give a positive answer to this question. In order to do that, we introduce a new class of operators: the confluence operators. These new operators can be useful in modelling negotiation processes.

1 Introduction

Belief change theory has produced a lot of different operators that models the different ways the beliefs of one (or some) agent(s) evolve over time. Among these operators, one can quote revision [1, 5, 10, 6], update [9, 8], merging [23, 14], abduction [20], extrapolation [4], etc.

In this paper we will focus on revision, update and merging. Let us first briefly describe these operators informally:

Revision Belief revision is the process of accomodating a new piece of evidence that is more reliable than the current beliefs of the agent. In belief revision the world is static, it is the beliefs of the agents that evolve.

Update In belief update the new piece of evidence denotes a change in the world. The world is dynamic, and these (observed) changes modify the beliefs of the agent.

Merging Belief merging is the process of defining the beliefs of a group of agents. So the question is: Given a set of agents that have their own beliefs, what can be considered as the beliefs of the group?

Apart from these intuitive differences between these operators, there are also close links between them. This is particularly clear when looking at the technical definitions. There are close relationship between revision [1, 5, 10] and KM update operators [9]. The first ones looking at the beliefs of the agents globally, the second ones looking at them locally (this sentence will be made formally clear later in the paper)[2]. There is also a close connection between revision and merging operators. In fact revision operators can be seen as particular cases of merging operators. From these two facts a very natural question arises: What is the family of operators that are a generalization of update operators in the same way merging operators generalize revision operators? Or, equivalently, what are the operators that can be considered as pointwise merging, just as KM update operators can be considered as pointwise belief revision. This can be outlined in the figure below. The aim of this

[1] This paper is a revised version of a paper that will be published in the Proceedings of the 11th European Conference on Logics in Artificial Intelligence (JELIA'08).

[2] See [8, 4, 15] for more discussions on update and its links with revision.

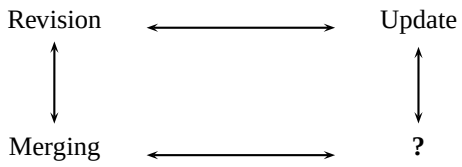


Figure 1: Revision - Update - Merging - Confluence

paper is to introduce and study the operators corresponding to the question mark. We will call these new operators confluence operators.

these new operators are more cautious than merging operators. This suggest that they can be used to define negotiation operators (see [2, 24, 22, 21, 12]), or as a first step of a negotiation process, in order to find all the possible negotiation results.

In order to illustrate the need for these new operators and also the difference of behaviour between merging and confluence we present the following small example.

Example 1 *Mary and Peter are planning to buy a car. Mary does not like a German car nor an expensive car. She likes small cars. Peter hesitates between a German, expensive but small car or a car which is not German, nor expensive and is a big car. Taking three propositional variables* `German_car, Expensive_car` *and* `Small_car` *in this order, Mary's desires are represented by* $mod(A) = \{001\}$ *and Peter's desires by* $mod(B) = \{111, 000\}$. *Most of the merging operators*[3] *give as solution (in semantical terms) the set* $\{001, 000\}$. *That is the same solution obtained when we suppose that Peter's desires are only a car which is not German nor expensive but a big car (*$mod(B') = \{000\}$*). The confluence operators will take into account the disjunctive nature of Peter's desires in a better manner and they will incorporate also the interpretations that are a trade-off between* 001 *and* 111. *For instance, the worlds* 011 *and* 101 *will be also in the solution if one use the confluence operator* $\diamond^{d_H, Gmax}$ *(defined in Section 7).*

This kind of operators is particularly adequate when the base describes a situation that is not perfectly known, or that can evolve in the future. For instance Peter's desires can either be imperfectly known (he wants one of the two situations but we do not know which one), or can evolve in the future (he will choose later between the two situations). In these situations the solutions proposed by confluence operators will be more adequate than the one proposed by merging operators. The solutions proposed by the confluence operators can be seen as all possible agreements in a negotiation process.

Belief merging is closely related to judgment aggregation as studied in political science and social choice theory (see e.g. [17, 18, 19]). An important difference is that in judgment aggregation there is an agenda on which the agents give their judgments. There is no such agenda in belief merging. The aim is to find the beliefs of the group. So this can be considered as a judgment aggregation problem where the agenda is the full set of formulae of the language (that are consistent with the integrity constraints). So, in a sense, judgment aggregation is an aggregation in a partial (incomplete) information framework (the only available information is about the formulae of the agenda), whereas belief merging is an aggregation in a complete (total/ideal) information framework.

Abstract negotiation processes have been studied both from belief merging [2, 24, 22, 21, 12] and judgment aggregation [16] perspectives. The definition of conciliation operators in this paper can be related to these works.

In the next section we will give the required definitions and notations. In Section 3 we will recall the postulates and representation theorems for revision, update, and merging, and state the links

[3] Such as $\triangle^{d_H, \Sigma}$ and $\triangle^{d_H, Gmax}$ [14].

between these operators. In Section 4 we define confluence operators. We provide a representation theorem for these operators in Section 5. In Section 6 we study the links between confluence operators and update and merging. In Section 7 we give examples of confluence operators. And we conclude in Section 8.

2 Preliminaries

We consider a propositional language $\mathcal{L}$ defined from a finite set of propositional variables $\mathcal{P}$ and the standard connectives, including $\top$ and $\bot$.

An interpretation ω is a total function from $\mathcal{P}$ to $\{0,1\}$. The set of all interpretations is denoted by $\mathcal{W}$. An interpretation ω is a model of a formula $\phi \in \mathcal{L}$ if and only if it makes it true in the usual truth functional way. $mod(\varphi)$ denotes the set of models of the formula φ, i.e., $mod(\varphi) = \{\omega \in \mathcal{W} \mid \omega \models \varphi\}$. When M is a set of models we denote by φ_M a formula such that $mod(\varphi_M) = M$.

A *base* K is a finite set of propositional formulae. In order to simplify the notations, in this work we will identify the base K with the formula φ which is the conjunction of the formulae of K[4].

A *profile* Ψ is a non-empty multi-set (bag) of bases $\Psi = \{\varphi_1, \ldots, \varphi_n\}$ (hence different agents are allowed to exhibit identical bases), and represents a group of n agents.

We denote by $\bigwedge \Psi$ the conjunction of bases of $\Psi = \{\varphi_1, \ldots, \varphi_n\}$, i.e., $\bigwedge \Psi = \varphi_1 \wedge \ldots \wedge \varphi_n$. A profile Ψ is said to be consistent if and only if $\bigwedge \Psi$ is consistent. The multi-set union is denoted by $\sqcup$.

A formula φ is complete if it has only one model. A profile Ψ is complete if all the bases of Ψ are complete formulae.

If $\leq$ denotes a pre-order on $\mathcal{W}$ (i.e., a reflexive and transitive relation), then $<$ denotes the associated strict order defined by $\omega < \omega'$ if and only if $\omega \leq \omega'$ and $\omega' \not\leq \omega$, and $\simeq$ denotes the associated equivalence relation defined by $\omega \simeq \omega'$ if and only if $\omega \leq \omega'$ and $\omega' \leq \omega$. A pre-order is *total* if $\forall \omega, \omega' \in \mathcal{W}$, $\omega \leq \omega'$ or $\omega' \leq \omega$. A pre-order that is not total is called *partial*. Let $\leq$ be a pre-order on A, and $B \subseteq A$, then $\min(B, \leq) = \{b \in B \mid \nexists a \in B \; a < b\}$.

3 Revision, Update and Merging

Let us now recall in this section some background on revision, update and merging, and their representation theorems in terms of pre-orders on interpretations. This will allow us to give the relationships between these operators.

3.1 Revision

Definition 1 (Katsuno-Mendelzon [10]) *An operator $\circ$ is an AGM belief revision operator if it satisfies the following properties:*

(R1) $\varphi \circ \mu \vdash \mu$

(R2) *If* $\varphi \wedge \mu \nvdash \bot$ *then* $\varphi \circ \mu \equiv \varphi \wedge \mu$

(R3) *If* $\mu \nvdash \bot$ *then* $\varphi \circ \mu \nvdash \bot$

(R4) *If* $\varphi_1 \equiv \varphi_2$ *and* $\mu_1 \equiv \mu_2$ *then* $\varphi_1 \circ \mu_1 \equiv \varphi_2 \circ \mu_2$

(R5) $(\varphi \circ \mu) \wedge \phi \vdash \varphi \circ (\mu \wedge \phi)$

(R6) *If* $(\varphi \circ \mu) \wedge \phi \nvdash \bot$ *then* $\varphi \circ (\mu \wedge \phi) \vdash (\varphi \circ \mu) \wedge \phi$

[4]Some approaches are sensitive to syntactical representation. In that case it is important to distinguish between K and the conjonction of its formulae (see e.g. [13]). But operators of this work are all syntax independant.

When one works with a finite propositional language the previous postulates, proposed by Katsuno and Mendelzon, are equivalent to AGM ones [1, 5]. In [10] Katsuno and Mendelzon give also a representation theorem for revision operators, showing that each revision operator corresponds to a faithful assignment, that associates to each base a plausibility preorder on interpretations (this idea can be traced back to Grove systems of spheres [7]).

Definition 2 *A faithful assignment is a function mapping each base φ to a pre-order $\leq_\varphi$ over interpretations such that:*

1. *If $\omega \models \varphi$ and $\omega' \models \varphi$, then $\omega \simeq_\varphi \omega'$*
2. *If $\omega \models \varphi$ and $\omega' \not\models \varphi$, then $\omega <_\varphi \omega'$*
3. *If $\varphi \equiv \varphi'$, then $\leq_\varphi = \leq_{\varphi'}$*

Theorem 1 (Katsuno-Mendelzon [10]) *An operator $\circ$ is a revision operator (ie. it satisfies (R1)-(R6)) if and only if there exists a faithful assignment that maps each base φ to a total pre-order $\leq_\varphi$ such that*

$$mod(\varphi \circ \mu) = \min(mod(\mu), \leq_\varphi).$$

This representation theorem is important because it provides a way to easily define revision operators by defining faithful assignments. But also because their are similar such theorems for update and merging (we will also show a similar result for confluence), and that these representations in term of assignments allow to more easily find links between these operators.

3.2 Update

Definition 3 (Katsuno-Mendelzon [9, 11]) *An operator $\diamond$ is a (partial) update operator if it satisfies the properties* **(U1)-(U8)**. *It is a total update operator if it satisfies the properties* **(U1)-(U5), (U8), (U9)**.

(U1) $\varphi \diamond \mu \vdash \mu$

(U2) *If $\varphi \vdash \mu$, then* $\varphi \diamond \mu \equiv \varphi$

(U3) *If $\varphi \nvdash \bot$ and $\mu \nvdash \bot$ then* $\varphi \diamond \mu \nvdash \bot$

(U4) *If $\varphi_1 \equiv \varphi_2$ and $\mu_1 \equiv \mu_2$ then* $\varphi_1 \diamond \mu_1 \equiv \varphi_2 \diamond \mu_2$

(U5) $(\varphi \diamond \mu) \wedge \phi \vdash \varphi \diamond (\mu \wedge \phi)$

(U6) *If $\varphi \diamond \mu_1 \vdash \mu_2$ and $\varphi \diamond \mu_2 \vdash \mu_1$, then* $\varphi \diamond \mu_1 \equiv \varphi \diamond \mu_2$

(U7) *If φ is a complete formula, then* $(\varphi \diamond \mu_1) \wedge (\varphi \diamond \mu_2) \vdash \varphi \diamond (\mu_1 \vee \mu_2)$

(U8) $(\varphi_1 \vee \varphi_2) \diamond \mu \equiv (\varphi_1 \diamond \mu) \vee (\varphi_2 \diamond \mu)$

(U9) *If φ is a complete formula and $(\varphi \diamond \mu) \wedge \phi \nvdash \bot$, then* $\varphi \diamond (\mu \wedge \phi) \vdash (\varphi \diamond \mu) \wedge \phi$

As for revision, there is a representation theorem in terms of faithful assignment.

Definition 4 *A faithful assignment is a function mapping each interpretation ω to a pre-order $\leq_\omega$ over interpretations such that if $\omega \neq \omega'$, then $\omega <_\omega \omega'$.*

One can easily check that this faithful assignment on interpretations is just a special case of the faithful assignment on bases defined in the previous section on the complete base corresponding to the interpretation.

Katsuno and Mendelzon give two representation theorems for update operators. The first representation theorem corresponds to partial pre-orders.

Theorem 2 (Katsuno-Mendelzon [9, 11]) *An update operator $\diamond$ satisfies (U1)-(U8) if and only if there exists a faithful assignment that maps each interpretation ω to a partial pre-order $\leq_\omega$ such that*

$$mod(\varphi \diamond \mu) = \bigcup_{\omega \models \varphi} \min(mod(\mu), \leq_{\varphi\{\omega\}})$$

And the second one corresponds to total pre-orders.

Theorem 3 (Katsuno-Mendelzon [9, 11]) *An update operator $\diamond$ satisfies (U1)-(U5), (U8) and (U9) if and only if there exists a faithful assignment that maps each interpretation ω to a total pre-order $\leq_\omega$ such that*

$$mod(\varphi \diamond \mu) = \bigcup_{\omega \models \varphi} \min(mod(\mu), \leq_{\varphi\{\omega\}})$$

3.3 Merging

Definition 5 (Konieczny-Pino Pérez [14]) *An operator $\triangle$ mapping a pair Ψ, μ (profile, formula) into a formula denoted $\triangle_\mu(\Psi)$ is an IC merging operator if it satisfies the following properties:*

(IC0) $\triangle_\mu(\Psi) \vdash \mu$

(IC1) *If μ is consistent, then $\triangle_\mu(\Psi)$ is consistent*

(IC2) *If $\bigwedge \Psi$ is consistent with μ, then $\triangle_\mu(\Psi) \equiv \bigwedge \Psi \wedge \mu$*

(IC3) *If $\Psi_1 \equiv \Psi_2$ and $\mu_1 \equiv \mu_2$, then $\triangle_{\mu_1}(\Psi_1) \equiv \triangle_{\mu_2}(\Psi_2)$*

(IC4) *If $\varphi_1 \vdash \mu$ and $\varphi_2 \vdash \mu$, then $\triangle_\mu(\{\varphi_1, \varphi_2\}) \wedge \varphi_1$ is consistent if and only if $\triangle_\mu(\{\varphi_1, \varphi_2\}) \wedge \varphi_2$ is consistent*

(IC5) $\triangle_\mu(\Psi_1) \wedge \triangle_\mu(\Psi_2) \vdash \triangle_\mu(\Psi_1 \sqcup \Psi_2)$

(IC6) *If $\triangle_\mu(\Psi_1) \wedge \triangle_\mu(\Psi_2)$ is consistent, then $\triangle_\mu(\Psi_1 \sqcup \Psi_2) \vdash \triangle_\mu(\Psi_1) \wedge \triangle_\mu(\Psi_2)$*

(IC7) $\triangle_{\mu_1}(\Psi) \wedge \mu_2 \vdash \triangle_{\mu_1 \wedge \mu_2}(\Psi)$

(IC8) *If $\triangle_{\mu_1}(\Psi) \wedge \mu_2$ is consistent, then $\triangle_{\mu_1 \wedge \mu_2}(\Psi) \vdash \triangle_{\mu_1}(\Psi)$*

There is also a representation theorem for merging operators in terms of pre-orders on interpretations [14].

Definition 6 *A* syncretic assignment *is a function mapping each profile Ψ to a total pre-order $\leq_\Psi$ over interpretations such that:*

1. *If $\omega \models \Psi$ and $\omega' \models \Psi$, then $\omega \simeq_\Psi \omega'$*
2. *If $\omega \models \Psi$ and $\omega' \not\models \Psi$, then $\omega <_\Psi \omega'$*
3. *If $\Psi_1 \equiv \Psi_2$, then $\leq_{\Psi_1} = \leq_{\Psi_2}$*
4. *$\forall \omega \models \varphi \; \exists \omega' \models \varphi' \; \omega' \leq_{\{\varphi\} \sqcup \{\varphi\}'} \omega$*

5. *If* $\omega \leq_{\Psi_1} \omega'$ *and* $\omega \leq_{\Psi_2} \omega'$*, then* $\omega \leq_{\Psi_1 \sqcup \Psi_2} \omega'$

6. *If* $\omega <_{\Psi_1} \omega'$ *and* $\omega \leq_{\Psi_2} \omega'$*, then* $\omega <_{\Psi_1 \sqcup \Psi_2} \omega'$

Theorem 4 (Konieczny-Pino Pérez [14]) *An operator* $\triangle$ *is an IC merging operator if and only if there exists a syncretic assignment that maps each profile* Ψ *to a total pre-order* $\leq_\Psi$ *such that*

$$mod(\triangle_\mu(\Psi)) = \min(mod(\mu), \leq_\Psi)$$

3.4 Revision vs Update

Intuitively revision operators bring a minimal change to the base by selecting the most plausible models among the models of the new information. Whereas update operators bring a minimal change to each possible world (model) of the base in order to take into account the change described by the new infomation whatever the possible world. So, if we look closely to the two representation theorems (propositions 1, 2 and 3), we easily find the following result:

Theorem 5 *If* $\circ$ *is a revision operator (i.e. it satisfies (R1)-(R6)), then the operator* $\diamond$ *defined by:*

$$\varphi \diamond \mu = \bigvee_{\omega \models \varphi} \varphi_{\{\omega\}} \circ \mu$$

is an update operator that satisfies (U1)-(U9).

Moreover, for each update operator $\diamond$*, there exists a revision operator* $\circ$ *such that the previous equation holds.*

As explained above this proposition states that update can be viewed as a kind of pointwise revision.

3.5 Revision vs Merging

Intuitively revision operators select in a formula (the new evidence) the closest information to a ground information (the old base). And, identically, IC merging operators select in a formula (the integrity constraints) the closest information to a ground information (a profile of bases).

So following this idea it is easy to make a correspondence between IC merging operators and belief revision operators [14]:

Theorem 6 (Konieczny-Pino Pérez [14]) *If* $\triangle$ *is an IC merging operator (it satisfies (IC0-IC8)), then the operator* $\circ$*, defined as* $\varphi \circ \mu = \triangle_\mu(\varphi)$*, is an AGM revision operator (it satisfies (R1-R6)).*

See [14] for more links between belief revision and merging.

4 Confluence operators

So now that we have made clear the connections sketched in figure 1 between revision, update and merging, let us turn now to the definition of confluence operators, that aim to be a pointwise merging, similarly as update is a pointwise revision, as explained in Section 3.4. Let us first define p-consistency for profiles.

Definition 7 *A profile* $\Psi = \{\varphi_1, \ldots \varphi_n\}$ *is* p-consistent *if all its bases are consistent, i.e* $\forall \varphi_i \in \Psi$*,* φ_i *is consistent.*

Note that p-consistency is much weaker than consistency, the former just asks that all the bases of the profile are consistent, while the later asks that the conjunction of all the bases is consistent.

Definition 8 *An operator $\diamond$ is a confluence operator if it satisfies the following properties:*

(UC0) $\diamond_\mu(\Psi) \vdash \mu$

(UC1) *If μ is consistent and Ψ is p-consistent, then $\diamond_\mu(\Psi)$ is consistent*

(UC2) *If Ψ is complete, Ψ is consistent and $\bigwedge \Psi \vdash \mu$, then $\diamond_\mu(\Psi) \equiv \bigwedge \Psi$*

(UC3) *If $\Psi_1 \equiv \Psi_2$ and $\mu_1 \equiv \mu_2$, then $\diamond_{\mu_1}(\Psi_1) \equiv \diamond_{\mu_2}(\Psi_2)$*

(UC4) *If φ_1 and φ_2 are complete formulae and $\varphi_1 \vdash \mu$, $\varphi_2 \vdash \mu$, then $\diamond_\mu(\{\varphi_1, \varphi_2\}) \wedge \varphi_1$ is consistent if and only $\diamond_\mu(\{\varphi_1, \varphi_2\}) \wedge \varphi_2$ is consistent*

(UC5) $\diamond_\mu(\Psi_1) \wedge \diamond_\mu(\Psi_2) \vdash \diamond_\mu(\Psi_1 \sqcup \Psi_2)$

(UC6) *If Ψ_1 and Ψ_2 are complete profiles and $\diamond_\mu(\Psi_1) \wedge \diamond_\mu(\Psi_2)$ is consistent, then $\diamond_\mu(\Psi_1 \sqcup \Psi_2) \vdash \diamond_\mu(\Psi_1) \wedge \diamond_\mu(\Psi_2)$*

(UC7) $\diamond_{\mu_1}(\Psi) \wedge \mu_2 \vdash \diamond_{\mu_1 \wedge \mu_2}(\Psi)$

(UC8) *If Ψ is a complete profile and if $\diamond_{\mu_1}(\Psi) \wedge \mu_2$ is consistent then $\diamond_{\mu_1 \wedge \mu_2}(\Psi) \vdash \diamond_{\mu_1}(\Psi) \wedge \mu_2$*

(UC9) $\diamond_\mu(\Psi \sqcup \{\varphi \vee \varphi'\}) \equiv \diamond_\mu(\Psi \sqcup \{\varphi\}) \vee \diamond_\mu(\Psi \sqcup \{\varphi'\})$

Some of the (UC) postulates are exactly the same as (IC) ones, just like some (U) postulates for update are exactly the same as (R) ones for revision.

In fact, (UC0), (UC3), (UC5) and (UC7) are exactly the same as the corresponding (IC) postulates. So the specificity of confluence operators lies in postulates (UC1), (UC2), (UC6), (UC8) and (UC9). (UC2), (UC4), (UC6) and (UC8) are close to the corresponding (IC) postulates, but hold for complete profiles only. The present formulation of (UC2) is quite similar to formulation of (U2) for update. Note that in the case of a complete profile the hypothesis of (UC2) is equivalent to ask coherence with the constraints, *i.e.* the hypothesis of (IC2). Postulates (UC8) and (UC9) are the main difference with merging postulates, and correspond also to the main difference between revision and KM update operators. (UC9) is the most important postulate, that defines confluence operators as pointwise agregation, just like (U8) defines update operators as pointwise revision. This will be expressed more formally in the next Section (Lemma 1).

5 Representation theorem for confluence operators

In order to state the representation theorem for confluence operators, we first have to be able to "localize" the problem. For update this is done by looking to each model of the base, instead of looking at the base (set of models) as a whole. So for "localizing" the aggregation process, we have to find what is the local view of a profile. That is what we call a state.

Definition 9 *A multi-set of interpretations will be called a* state. *We use the letter e, possibly with subscripts, for denoting states. If $\Psi = \{\varphi_1, \ldots, \varphi_n\}$ is a profile and $e = \{\omega_1, \ldots, \omega_n\}$ is a state such that $\omega_i \models \varphi_i$ for each i, we say that e is a state of the profile Ψ, or that the state e models the profile Ψ, that will be denoted by $e \models \Psi$. If $e = \{\omega_1, \ldots, \omega_n\}$ is a state, we define the profile Ψ_e by putting $\Psi_e = \{\varphi_{\{\omega_1\}}, \ldots, \varphi_{\{\omega_n\}}\}$.*

State is an interesting notion. If we consider each base as the current point of view (goals) of the corresponding agent (that can be possibly strengthened in the future) then states are all possible negotiation starting points.

States are the points of interest for confluence operators (like interpretations are for update), as stated in the following Lemma:

Lemma 1 *If $\diamond$ satisfies (UC3) and (UC9) then $\diamond$ satisfies the following*

$$\diamond_\mu(\Psi) \equiv \bigvee_{e \models \Psi} \diamond_\mu(\Psi_e)$$

Defining profile entailment by putting $\Psi \vdash \Psi'$ iff every state of Ψ is a state of Ψ', the previous Lemma has as a corollary the following:

Corollary 1 *If $\diamond$ is a confluence operator then it is monotonic in the profiles, that means that if $\Psi \vdash \Psi'$ then $\diamond_\mu(\Psi) \vdash \diamond_\mu(\Psi')$*

This monotony property, that is not true in the case of merging operators, shows one of the big differences between merging and confluence operators. Remark that there is a corresponding monotony property for update.

Like revision's faithful assignments that have to be "localized" to interpretations for update, merging's syncretic assignments have to be localized to states for confluence.

Definition 10 *A* distributed assignment *is a function mapping each state e to a total pre-order $\leq_e$ over interpretations such that:*

1. $\omega <_{\{\omega,\ldots,\omega\}} \omega'$ *if* $\omega' \neq \omega$
2. $\omega \simeq_{\{\omega,\omega'\}} \omega'$
3. *If* $\omega \leq_{e_1} \omega'$ *and* $\omega \leq_{e_2} \omega'$, *then* $\omega \leq_{e_1 \sqcup e_2} \omega'$
4. *If* $\omega <_{e_1} \omega'$ *and* $\omega \leq_{e_2} \omega'$, *then* $\omega <_{e_1 \sqcup e_2} \omega'$

Now we can state the main result of this paper, that is the representation theorem for confluence operators.

Theorem 7 *An operator $\diamond$ is a confluence operator if and only if there exists a distributed assignment that maps each state e to a total pre-order $\leq_e$ such that*

$$mod(\diamond_\mu(\Psi)) \quad = \quad \bigcup_{e \models \Psi} \min(mod(\mu), \leq_e) \tag{1}$$

Unfortunately, we have to omit the proof for space reasons. Nevertheless, we indicate the most important ideas therein. As it is usual, the *if* condition is done by checking each property without any major difficulty. In order to verify the *only if* condition we have to define a distributed assigment. This is done in the following way: for each state e we define a total pre-order $\leq_e$ by putting $\forall \omega, \omega' \in \mathcal{W}$ $\omega \leq_e \omega'$ if and only if $\omega \models \diamond_{\varphi_{\{\omega,\omega'\}}}(\Psi_e)$. Then, the main difficulties are to prove that this is indeed a distributed assigment and that the equation (1) holds. In particular, Lema 1 is very helpful for proving this last equation.

Note that this theorem is still true if we remove respectively the postulate (UC4) from the required postulates for confluence operators and the condition 2 from distributed assignments.

6 Confluence vs Update and Merging

So now we are able to state the proposition that shows that update is a special case of confluence, just as revision is a special case of merging.

Theorem 8 *If $\diamond$ is a confluence operator (i.e. it satisfies (UC0-UC9)), then the operator $\diamond$, defined as $\varphi \diamond \mu = \diamond_\mu(\varphi)$, is an update operator (i.e. it satisfies (U1-U9)).*

Concerning merging operators, one can see easily that the restriction of a syncretic assignment to a complete profile is a distributed assignment. From that we obtain the following result (the one corresponding to Theorem 5):

Theorem 9 *If $\triangle$ is an IC merging operator (i.e. it satisfies (IC0-IC8)) then the operator $\diamond$ defined by*

$$\diamond_\mu(\Psi) = \bigvee_{e \models \Psi} \triangle_\mu(\Psi_e)$$

is a confluence operator (i.e. it satisfies (UC0-UC9)).

Moreover, for each confluence operator $\diamond$, there exists a merging operator $\triangle$ such that the previous equation holds.

It is interesting to note that this theorem shows that every merging operator can be used to define a confluence operator, and explains why we can consider confluence as a pointwise merging.

Unlike Theorem 5, the second part of the previous theorem doesn't follow straightforwardly from the representation theorems. We need to build a syncretic assignment extending the distributed assignment representing the confluence operator. In order to do that we can use the following construction: Each pre-order $\leq_e$ defines naturally a rank function r_e on natural numbers. Then we put

$$\omega \leq_\Psi \omega' \quad \text{if and only if} \quad \sum_{e \models \Psi} r_e(\omega) \leq \sum_{e \models \Psi} r_e(\omega')$$

As a corollary of the representation theorem we obtain the following

Corollary 2 *If $\diamond$ is a confluence operator then the following property holds:*

If $\bigwedge \Psi \vdash \mu$ and Ψ is consistent then $\bigwedge \Psi \wedge \mu \vdash \diamond_\mu(\Psi)$

But unlike merging operators, we don't have generally $\diamond_\mu(\Psi) \vdash \bigwedge \Psi \wedge \mu$.

Note that this "half of (IC2)" property is similar to the "half of (R2)" satisfied by update operators.

This corollary is interesting since it underlines an important difference between merging and confluence operators. If all the bases agree (i.e. if their conjunction is consistent), then a merging operator gives as result exactly the conjunction, whereas a confluence operator will give this conjunction plus additional results. This is useful if the bases do not represent interpretations that are considered equivalent by the agent, but uncertain information about the agent's current or future state of mind.

7 Example

In this section we will illustrate the behaviour of confluence operators on an example. We can define confluence operators very similarly to merging operators, by using a distance and an aggregation function.

Definition 11 *A* pseudo-distance *between interpretations is a total function $d : \mathcal{W} \times \mathcal{W} \mapsto \mathbb{R}^+$ s.t. for any $\omega, \omega' \in \mathcal{W}$: $d(\omega, \omega') = d(\omega', \omega)$, and $d(\omega, \omega') = 0$ if and only if $\omega = \omega'$.*

A widely used distance between interpretations is the Dalal distance [3], denoted d_H, that is the Hamming distance between interpretations (the number of propositional atoms on which the two interpretations differ).

Definition 12 *An* aggregation function f *is a total function associating a nonnegative real number to every finite tuple of nonnegative real numbers s.t. for any* $x_1, \ldots, x_n, x, y \in \mathbb{R}^+$*:*

- *if* $x \leq y$*, then* $f(x_1, \ldots, x, \ldots, x_n) \leq f(x_1, \ldots, y, \ldots, x_n)$ *(non-decreasingness)*
- $f(x_1, \ldots, x_n) = 0$ *if and only if* $x_1 = \ldots = x_n = 0$ *(minimality)*
- $f(x) = x$ *(identity)*

Sensible aggregation functions are for instance $\max$, sum, or leximax ($Gmax$) [5] [14].

Definition 13 (distance-based confluence operators) *Let* d *be a pseudo-distance between interpretations and* f *be an aggregation function. The result* $\diamond_\mu^{d,f}(\Psi)$ *of the confluence of* Ψ *given the integrity constraints* μ *is defined by:* $mod(\diamond_\mu^{d,f}(\Psi)) = \bigcup_{e \models \Psi} \min(mod(\mu), \leq_e)$*, where the pre-order* $\leq_e$ *on* $\mathcal{W}$ *induced by* e *is defined by:*

- $\omega \leq_e \omega'$ *if and only if* $d(\omega, e) \leq d(\omega', e)$*, where*
- $d(\omega, e) = f(d(\omega, \omega_1) \ldots, d(\omega, \omega_n))$ *with* $e = \{\omega_1, \ldots, \omega_n\}$.

It is easy to check that by using usual aggregation functions we obtain confluence operators.

Proposition 1 *Let* d *be any distance,* $\diamond_\mu^{d,\Sigma}(\Psi)$ *and* $\diamond_\mu^{d,Gmax}(\Psi)$ *are confluence operators (i.e. they satisfy (UC0)-(UC9)).*

Example 2 *Let us consider a profile* $\Psi = \{\varphi_1, \varphi_2, \varphi_3, \varphi_4\}$ *and an integrity constraint* μ *defined on a propositional language built over four symbols, as follows:* $mod(\mu) = \mathcal{W} \setminus \{0110, 1010, 1100, 1110\}$*,* $mod(\varphi_1) = mod(\varphi_2) = \{1111, 1110\}$*,* $mod(\varphi_3) = \{0000\}$*, and* $mod(\varphi_4) = \{1110, 0110\}$.

$\mathcal{W}$	1111	1110	0000	0110	e_1		e_2		e_3		e_4		e_5		e_6		$\diamond_\mu^{d,\Sigma}$	$\diamond_\mu^{d,\text{Gmax}}$
					Σ	*Gmax*	Σ	*Gmax*	Σ	*Gmax*	Σ	*Gmax*	Σ	*Gmax*	Σ	*Gmax*		
0000	*4*	*3*	*0*	*2*	*11*	*4430*	*10*	*4420*	*10*	*4330*	*9*	*4320*	*9*	*3330*	*8*	*3320*		
0001	*3*	*4*	*1*	*3*	*11*	*4331*	*10*	*3331*	*12*	*4431*	*11*	*4331*	*13*	*4441*	*12*	*4431*		
0010	*3*	*2*	*1*	*1*	*9*	*3321*	*8*	*3311*	*8*	*3221*	*7*	**3211**	**7**	**2221**	**6**	**2211**	×	×
0011	*2*	*3*	*2*	*2*	*9*	*3222*	*8*	**2222**	*10*	*3322*	*9*	*3222*	*11*	*3332*	*10*	*3322*		×
0100	*3*	*2*	*1*	*1*	*9*	*3321*	*8*	*3311*	*8*	**3221**	**7**	**3211**	**7**	**2221**	**6**	**2211**	×	×
0101	*2*	*3*	*2*	*2*	*9*	*3222*	*8*	**2222**	*10*	*3322*	*9*	*3222*	*11*	*3332*	*10*	*3322*		×
0110	*2*	*1*	*2*	*0*	*7*	*2221*	*6*	*2220*	*6*	*2211*	*5*	*2210*	*5*	*2111*	*4*	*2110*		
0111	*1*	*2*	*3*	*1*	*7*	**3211**	**6**	*3111*	*8*	**3221**	**7**	**3211**	*9*	*3222*	*8*	*3221*	×	×
1000	*3*	*2*	*1*	*3*	*9*	*3321*	*10*	*3331*	*8*	**3221**	*9*	*3321*	**7**	**2221**	*8*	*3221*	×	×
1001	*2*	*3*	*2*	*4*	*9*	*3222*	*10*	*4222*	*10*	*3322*	*11*	*4322*	*11*	*3332*	*12*	*4332*		
1010	*2*	*1*	*2*	*2*	*7*	*2221*	*8*	*2222*	*6*	*2211*	*7*	*2221*	*5*	*2111*	*6*	*2211*		
1011	*1*	*2*	*3*	*3*	*7*	**3211**	*8*	*3311*	*8*	**3221**	*9*	*3321*	*9*	*3222*	*10*	*3322*		×
1100	*2*	*1*	*2*	*2*	*7*	*2221*	*8*	*2222*	*6*	*2211*	*7*	*2221*	*5*	*2111*	*6*	*2211*		
1101	*1*	*2*	*3*	*3*	*7*	**3211**	*8*	*3311*	*8*	**3221**	*9*	*3321*	*9*	*3222*	*10*	*3322*		×
1110	*1*	*0*	*3*	*1*	*5*	*3110*	*6*	*3111*	*4*	*3100*	*5*	*3110*	*3*	*3000*	*4*	*3100*		
1111	*0*	*1*	*4*	*2*	**5**	*4100*	**6**	*4200*	**6**	*4110*	**7**	*4210*	**7**	*4111*	*8*	*4211*	×	

Table 1: Computations of $mod(\diamond_\mu^{d,\Sigma}(\Psi))$ and $mod(\diamond_\mu^{d,Gmax}(\Psi))$

The computations are reported in Table 1. The shadowed lines correspond to the interpretations rejected by the integrity constraints. Thus the result has to be taken among the interpretations that are not shadowed. The states that model the profile are the following ones:
$e_1 = \{1111, 1111, 0000, 1110\}, e_2 = \{1111, 1111, 0000, 0110\},$
$e_3 = \{1111, 1110, 0000, 1110\}, e_4 = \{1110, 1111, 0000, 0110\},$
$e_5 = \{1110, 1110, 0000, 1110\}, e_6 = \{1110, 1110, 0000, 0110\}.$

[5] leximax ($Gmax$) is usually defined using lexicographic sequences, but it can be easily represented by reals to fit the above definition (see e.g. [13]).

For each state, the Table gives the distance between the interpretation and this state for the Σ *aggregation function, and for the* $Gmax$ *one. So one can then look at the best interpretations for each state.*

So for instance for $\diamond_\mu^{d,\Sigma}(\Psi)$*,* e_1 *selects the interpretation* 1111*,* e_2 *selects* 0111 *and* 1111*, etc. So, taking the union of the interpretations selected by each state, gives* $mod(\diamond_\mu^{d,\Sigma}(\Psi)) = \{0010, 0100, 0111, 1000, 1111\}$.

Similarly we obtain $mod(\diamond_\mu^{d,Gmax}(\Psi)) = \{0100, 0011, 0010, 0101, 0111, 1000, 1011, 1101\}$.

8 Conclusion

We have proposed in this paper a new family of change operators. Confluence operators are pointwise merging, just as update can be seen as a pointwise revision. We provide an axiomatic definition of this family, a representation theorem in terms of pre-orders on interpretations, and provide examples of these operators.

In this paper we define confluence operators as generalization to multiple bases of total update operators (i.e. which semantical counterpart are total pre-orders). A perspective of this work is to try to extend the result to partial update operators.

As Example 1 suggests, these operator can prove meaningful to aggregate the goals of a group of agents. They seem to be less adequate for aggregating beliefs, where the global minimization done by merging operators is more appropriate for finding the most plausible worlds. This distinction between goal and belief aggregation is a very interesting perspective, since, as far as we know, no such axiomatic distinction as been ever discussed.

Acknowledgements

The idea of this paper comes from discussions in the 2005 and 2007 Dagstuhl seminars ($\#5321$ and $\#7351$) *"Belief Change in Rational Agents"* . The authors would like to thank the Schloss Dagstuhl institution, and the participants of the seminars, especially Andreas Herzig for the initial question *"If merging can be seen as a generalization of revision, what is the generalization of update ?"*. Here is an answer !

The second author was partially supported by a research grant of the Mairie de Paris and by the project CDCHT-ULA N° C-1451-07-05-A. Part of this work was done when the second author was a visiting professor at CRIL (CNRS UMR 8188) from September to December 2007 and a visiting researcher at TSI Department of Telecom ParisTech (CNRS UMR 5141 LTCI) from January to April 2008. The second author thanks to CRIL and TSI Department for the excellent working conditions.

References

[1] C. E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

[2] R. Booth. Social contraction and belief negotiation. In *Proceedings of the Eighth Conference on Principles of Knowledge Representation and Reasoning (KR'02)*, pages 374–384, 2002.

[3] M. Dalal. Investigations into a theory of knowledge base revision: preliminary report. In *Proceedings of the American National Conference on Artificial Intelligence (AAAI'88)*, pages 475–479, 1988.

[4] F. Dupin de Saint-Cyr and J. Lang. Belief extrapolation (or how to reason about observations and unpredicted change). *Proceedings of the Eighth Conference on Principles of Knowledge Representation and Reasoning (KR'02)*, pages 497–508, 2002.

[5] P. Gärdenfors. *Knowledge in flux*. MIT Press, 1988.

[6] P. Gärdenfors, editor. *Belief Revision*. Cambridge University Press, 1992.

[7] A. Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17(157-180), 1988.

[8] A. Herzig and O. Rifi. Update operations: a review. In *Proceedings of the Thirteenth European Conference on Artificial Intelligence (ECAI'98)*, pages 13–17, 1998.

[9] H. Katsuno and A. O. Mendelzon. On the difference between updating a knowledge base and revising it. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 387–394, 1991.

[10] H. Katsuno and A. O. Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52:263–294, 1991.

[11] H. Katsuno and A. O. Mendelzon. On the difference between updating a knowledge base and revising it. In *Belief revision*, volume 29 of *Cambridge Tracts Theoret. Comput. Sci.*, pages 183–203. Cambridge Univ. Press, Cambridge, 1992.

[12] S. Konieczny. Belief base merging as a game. *Journal of Applied Non-Classical Logics*, 14(3):275–294, 2004.

[13] S. Konieczny, J. Lang, and P. Marquis. DA^2 merging operators. *Artificial Intelligence*, 157(1-2):49–79, 2004.

[14] S. Konieczny and R. Pino Pérez. Merging information under constraints: a logical framework. *Journal of Logic and Computation*, 12(5):773–808, 2002.

[15] J. Lang. Belief update revisited. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'07)*, pages 2517–2522, 2007.

[16] C. List. Group deliberation and the revision of judgments: An impossibility result.

[17] C. List and P. Pettit. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy*, 18:89–110, 2002.

[18] C. List and P. Pettit. Aggregating sets of judgments: Two impossibility results compared. *Synthese*, 140(1-2):207–235, 2004.

[19] C. List and C. Puppe. Judgment aggregation: a survey. In C. Puppe P. Anand and P. Pattaniak, editors, *Oxford Handbook of Rational and Social Choice*. Oxford University Press, 2009.

[20] J. Lobo and C. Uzcátegui. Abductive change operators. *Fundamenta Informaticae*, 27(4):385–411, 1996.

[21] T. Meyer, N. Foo, D. Zhang, and R. Kwok. Logical foundations of negotiation: Outcome, concession and adaptation. In *Proceedings of the American National Conference on Artificial Intelligence (AAAI'04)*, pages 293–298, 2004.

[22] T. Meyer, N. Foo, D. Zhang, and R. Kwok. Logical foundations of negotiation: Strategies and preferences. In *Proceedings of the Ninth Conference on Principles of Knowledge Representation and Reasoning (KR'04)*, pages 311–318, 2004.

[23] P. Z. Revesz. On the semantics of arbitration. *International Journal of Algebra and Computation*, 7(2):133–160, 1997.

[24] D. Zhang, N. Foo, T. Meyer, and R. Kwok. Negotiation as mutual belief revision. In *Proceedings of the American National Conference on Artificial Intelligence (AAAI'04)*, pages 317–322, 2004.

Welfare properties of argumentation-based semantics[1]

Kate Larson and Iyad Rahwan

Abstract

Since its introduction in the mid-nineties, Dung's theory of abstract argumentation frameworks has been influential in artificial intelligence. Dung viewed arguments as abstract entities with a binary defeat relation among them. This enabled extensive analysis of different (semantic) argument acceptance criteria. However, little attention has been given to comparing such criteria in relation to the preferences of self-interested agents who may have conflicting preferences over the final status of arguments. In this paper, we define a number of agent preference relations over argumentation outcomes. We then analyse different argument evaluation rules taking into account the preferences of individual agents.

1 Introduction

Negotiation is at the core of multiagent systems since it provides procedures so that agents can find beneficial agreements. While approaches based on game-theory have proved to be highly influential [9], an alternative approach for conducting negotiations is through argumentation [8]. In argumentation, the focus is on how assertions or statements are proposed and resolved in settings where agents may have different opinions and goals. Dung presented one of the most influential computational models of argument [6]. Arguments are viewed as abstract entities, with a binary defeat relation among them. This view of argumentation enables high-level analysis while abstracting away from the internal structure of individual arguments. In Dung's approach, given a set of arguments and a binary defeat relation, a rule specifies which arguments should be accepted. A variety of such rules have been analysed using intuitive *objective* logical criteria such as consistency or self-defence [2].

Most research that employs Dung's approach discounts the fact that argumentation takes place among self-interested agents, who may have conflicting preferences over which arguments end up being accepted, rejected, or undecided. As such, argumentation can (and arguably should) be studied as an economic mechanism in which determining the acceptability status of arguments is akin to allocating resources.

In any allocation mechanism involving multiple agents (be it resource allocation or argument status assignment), two complementary issues are usually studied. On one hand, we may analyse the agents' incentives in order to predict the equilibrium outcome of rational strategies. On the other hand, we may analyse the properties of the outcomes themselves in order to compare different allocation mechanisms. The above issues are the subject of study of the field of game theory and welfare economics, respectively.

The study of incentives in abstract argumentation has commenced recently [7]. To complement this work, in this paper we initiate the study of *preference* and *welfare* in abstract argumentation mechanisms. To this end, we define several new classes of agent preferences over the outcomes of an argumentation process. We then analyse different existing rules for argument status assignment in terms of how they satisfy the preferences of the agents involved. Our focus in this paper is on the property of Pareto optimality, which measures whether an outcome can be improved for one agent without harming other agents. We also discuss more refined social welfare measures.

The paper makes two distinct contributions to the state-of-the-art in computational models of argument. First, the paper extends Rahwan and Larson's definition of argumentation outcomes [7]

[1] A preliminary version of this paper appeared in AAAI 2008, with the title *Pareto optimality in abstract argumentation*.

to account for complete labellings of arguments (as opposed to accepted arguments only). This allows us to define a number of novel preference criteria that arguing agents may have.

The second contribution of this paper is the comparison of different argumentation semantics using a well-known social welfare measure, namely Pareto optimality. To our knowledge, this is the first attempt to evaluate Dung semantics in terms of the social desirability of its outcomes. In particular, we show that in many cases, these semantics fail to fully characterise Pareto optimal outcomes. Thus, when the semantics provides multiple possible argument status assignments, our analysis presents a new criterion for selecting among those.

2 Background

In this section, we briefly outline key elements of abstract argumentation frameworks. We begin with Dung's abstract characterisation of an argumentation system [6]:

Definition 1 (Argumentation framework)**.** *An* argumentation framework *is a pair* $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ *where* $\mathcal{A}$ *is a set of arguments and* $\rightharpoonup \subseteq \mathcal{A} \times \mathcal{A}$ *is a defeat relation. We say that an argument* α defeats *an argument* β *if* $(\alpha, \beta) \in \rightharpoonup$ *(sometimes written* $\alpha \rightharpoonup \beta$*).*[2]

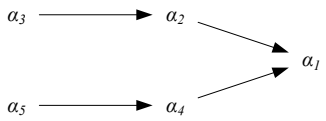


Figure 1: A simple argument graph

An argumentation framework can be represented as a directed graph in which vertices are arguments and directed arcs characterise defeat among arguments. An example argument graph is shown in Figure 1. Argument α_1 has two defeaters (*i.e.* counter-arguments) α_2 and α_4, which are themselves defeated by arguments α_3 and α_5 respectively.

Let $S^+ = \{\beta \in \mathcal{A} \mid \alpha \rightharpoonup \beta \text{ for some } \alpha \in S\}$. Also let $\alpha^- = \{\beta \in \mathcal{A} \mid \beta \rightharpoonup \alpha\}$. We first characterise the fundamental notions of conflict-free and defence.

Definition 2 (Conflict-free, Defence)**.** *Let* $\langle \mathcal{A}, \rightharpoonup \rangle$ *be an argumentation framework and let* $S \subseteq \mathcal{A}$ *and let* $\alpha \in \mathcal{A}$*.*

- S *is* conflict-free *if* $S \cap S^+ = \emptyset$*.*
- S defends *argument* α *if* $\alpha^- \subseteq S^+$*. We also say that argument* α *is* acceptable with respect to S*.*

Intuitively, a set of arguments is *conflict free* if no argument in that set defeats another. A set of arguments *defends* a given argument if it defeats all its defeaters. In Figure 1, for example, $\{\alpha_3, \alpha_5\}$ defends α_1. We now look at different semantics that characterise the *collective acceptability* of a set of arguments.

Definition 3 (Characteristic function)**.** *Let* $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ *be an argumentation framework. The* characteristic function *of* AF *is* $\mathcal{F}_{AF}$*:* $2^{\mathcal{A}} \rightarrow 2^{\mathcal{A}}$ *such that, given* $S \subseteq \mathcal{A}$*, we have* $\mathcal{F}_{AF}(S) = \{\alpha \in \mathcal{A} \mid S \text{ defends } \alpha\}$*.*

When there is no ambiguity about the argumentation framework in question, we will use $\mathcal{F}$ instead of $\mathcal{F}_{AF}$.

[2] We restrict ourselves to finite sets of arguments.

Definition 4 (Acceptability semantics)**.** *Let S be a conflict-free set of arguments in framework $\langle \mathcal{A}, \rightharpoonup \rangle$.*

- *S is* admissible *if it is conflict-free and defends every element in S (i.e. if $S \subseteq \mathcal{F}(S)$).*
- *S is a* complete extension *if $S = \mathcal{F}(S)$.*
- *S is a* grounded extension *if it is the minimal (w.r.t. set-inclusion) complete extension (or, alternatively, if S is the least fixed-point of $\mathcal{F}(.)$).*
- *S is a* preferred extension *if it is a maximal (w.r.t. set-inclusion) complete extension (or, alternatively, if S is a maximal admissible set).*
- *S is a* stable extension *if $S^+ = \mathcal{A} \backslash S$.*
- *S is a* semi-stable extension *if S is a complete extension of which $S \cup S^+$ is maximal.*

Intuitively, a set of arguments is *admissible* if it is a conflict-free set that defends itself against any defeater – in other words, if it is a conflict free set in which each argument is acceptable with respect to the set itself.

An admissible set S is a *complete extension* if and only if *all* arguments defended by S are also in S (that is, if S is a fixed point of the operator $\mathcal{F}$). There may be more than one complete extension, each corresponding to a particular consistent and self-defending viewpoint.

A *grounded extension* contains all the arguments which are not defeated, as well as the arguments which are defended directly or indirectly by non-defeated arguments. This can be seen as a non-committal view (characterised by the *least* fixed point of $\mathcal{F}$). As such, there always exists a unique grounded extension. Dung [6] showed that in finite argumentation systems, the grounded extension can be obtained by an iterative application of the characteristic function to the empty set. For example, in Figure 1 the grounded extension is $\{\alpha_1, \alpha_3, \alpha_5\}$, which is the only complete extension.

A *preferred extension* is a bolder, more committed position that cannot be extended – by accepting more arguments – without causing inconsistency. Thus a preferred extension can be thought of as a maximal consistent set of hypotheses. There may be multiple preferred extensions, and the grounded extension is included in all of them.

Finally, a set of arguments is a *stable extension* if it is a preferred extension that defeats every argument which does not belong to it. A *semi-stable extension* requires the weaker condition that the set of arguments defeated is maximal.

Crucial to our subsequent analysis is the notion of *argument labelling* [3], which specifies a particular *outcome* of argumentation. It specifies which arguments are accepted (labelled `in`), which ones are rejected (labelled `out`), and which ones whose acceptance or rejection could not be decided (labelled `undec`). Labellings must satisfy the condition that an argument is `in` if and only if all of its defeaters are `out`. An argument is `out` if and only if at least one of its defeaters is `in`.

Definition 5 (Argument Labelling)**.** *Let $\langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. An* argument labelling *is a total function $L : \mathcal{A} \rightarrow \{\texttt{in}, \texttt{out}, \texttt{undec}\}$ such that:*

- $\forall \alpha \in \mathcal{A} \ : \ (L(\alpha) = \texttt{out} \equiv \exists \beta \in \mathcal{A}$ *such that* $(\beta \rightharpoonup \alpha$ *and* $L(\beta) = \texttt{in}))$*; and*
- $\forall \alpha \in \mathcal{A} \ : \ (L(\alpha) = \texttt{in} \equiv \forall \beta \in \mathcal{A} \ : \ (\ $*if* $\beta \rightharpoonup \alpha$ *then* $L(\beta) = \texttt{out}))$

We will make use of the following notation.

Definition 6. *Let $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework, and L a labelling over AF. We define:*

- $\texttt{in}(L) = \{\alpha \in \mathcal{A} \mid L(\alpha) = \texttt{in}\}$

- $\texttt{out}(L) = \{\alpha \in \mathcal{A} \mid L(\alpha) = \texttt{out}\}$
- $\texttt{undec}(L) = \{\alpha \in \mathcal{A} \mid L(\alpha) = \texttt{undec}\}$

In the rest of the paper, by slight abuse of notation, when we refer to a labelling L as an *extension*, we will be referring to the set of accepted arguments $\texttt{in}(L)$.

Caminada [3] established a correspondence between properties of labellings and the different extensions. These are summarised in Table 1.

Extensions	Restrictions on Labellings
complete	all labellings
grounded	minimal `in` minimal `out` maximal `undec`
preferred	maximal `in` maximal `out`
semi-stable	minimal `undec`
stable	empty `undec`

Table 1: The relationships between extensions and labellings.

3 Agent Preferences

Abstract argumentation frameworks have typically been analysed without taking into account the agents involved. This is because the focus has mostly been on studying the logically intuitive properties of argument acceptance criteria [2]. Recently research has commenced on evaluating argument acceptance criteria taking into account agents' strategic behaviour [7]. In this paper, we focus on developing an understanding of the underlying preferences of the agents and how these can be used in refining outcomes of the argumentation process. While we assume that agents are non-strategic, this paper complements our earlier work in that strategic behaviour is often motivated by underlying preferences.

In this paper we view an outcome as an *argument labelling*, specifying not only which arguments are accepted, but also which ones are rejected or undecided. Thus the set $\mathcal{L}$ of possible outcomes is exactly the set of all possible labellings of all arguments.

We let $\theta_i \in \Theta_i$ denote the *type* of agent $i \in I$ which is drawn from some set of possible types Θ_i. The type represents the private information and preferences of the agent. More precisely, θ_i determines the set $\mathcal{A}_i$ of arguments available to agent i, as well as the preference criterion used to evaluate outcomes. We place no restrictions on the argument sets of agents, and for $i \neq j$ it is possible that $\mathcal{A}_i \cap \mathcal{A}_j \neq \emptyset$. An agent's preferences are over *outcomes* $L \in \mathcal{L}$. By $L_1 \succeq_i L_2$ we denote that agent i *weakly prefers* (or simply *prefers*) outcome L_1 to L_2. We say that agent i *strictly prefers* outcome L_1 to L_2, written $L_1 \succ_i L_2$, if and only if $L_1 \succeq_i L_2$ but not $L_2 \succeq_i L_1$. Finally, we say that agent i is *indifferent* between outcomes L_1 and L_2, written $L_1 \sim_i L_2$, if and only if both $L_1 \succeq_i L_2$ and $L_2 \succeq_i L_1$.

We define agents' preferences with respect to restricted sets of arguments in order to model situations where agents have potentially different *domains of knowledge*. As a motivating example, consider a court case where a medical expert is called as an expert witness. This expert can put forward arguments related to medical forensics, but would be unable to comment on legal issues. Similarly, an agent's arguments can be limited by their *position of knowledge*. For example, a friend may be in a position to comment on someone's character, while a stranger's comments would not be of interest.

While many classes of preferences are possible, in this paper we focus on *self-interested* preferences. By this we mean that we are interested in preference structures where each agent i is only

interested in the status (*i.e.* labelling) of its own arguments and not on the particular status of other agents' arguments. We also emphasize that we assume that all agents understand and share the underlying argumentation system. Thus, the question of merging argumentation systems is outside the scope of this paper [5].

We start with *individual acceptability maximising preferences* [7]. Under these preferences, each agent wants to maximise the number of arguments in $\mathcal{A}_i$ that end up being accepted.

Definition 7 (Acceptability maximising preferences)**.** *An agent i has* individual acceptability maximising preferences *if $\forall L_1, L_2 \in \mathcal{L}$ such that $|\texttt{in}(L_1) \cap \mathcal{A}_i| \geq |\texttt{in}(L_2) \cap \mathcal{A}_i|$, we have $L_1 \succeq_i L_2$.*

An agent may, instead, aim to minimise the number of arguments in $\mathcal{A}_i$ that end up rejected.

Definition 8 (Rejection minimising preferences)**.** *An agent i has* individual rejection minimising preferences *if $\forall L_1, L_2 \in \mathcal{L}$ such that $|\texttt{out}(L_1) \cap \mathcal{A}_i| \leq |\texttt{out}(L_2) \cap \mathcal{A}_i|$, we have $L_1 \succeq_i L_2$.*

An agent may prefer outcomes which minimise uncertainty by having as few undecided arguments as possible.

Definition 9 (Decisive preferences)**.** *An agent i has decisive preferences if $\forall L_1, L_2 \in \mathcal{L}$ if $|\texttt{undec}(L_1) \cap \mathcal{A}_i| \leq |\texttt{undec}(L_2) \cap \mathcal{A}_i|$ then $L_1 \succeq_i L_2$.*

An agent may only be interested in getting *all* of its arguments collectively accepted.

Definition 10 (All-or-nothing preferences)**.** *An agent i has* all-or-nothing preferences *if and only if $\forall L_1, L_2 \in \mathcal{L}$, if $\mathcal{A}_i \subseteq \texttt{in}(L_1)$ and $\mathcal{A}_i \nsubseteq \texttt{in}(L_2)$, then $L_1 \succ_i L_2$, otherwise $L_1 \sim_i L_2$.*

Instead of having all of its arguments collectively accepted, an agent may be interested in having one particular *focal* argument accepted.

Definition 11 (Focal-argument preferences)**.** *An agent i has* focal-argument preferences *if and only if there exists some argument $\alpha_i^* \in \mathcal{A}_i$ such that $\forall L_1, L_2 \in \mathcal{L}$ if $\alpha_i^* \in \texttt{in}(L_1)$ and $\alpha_i^* \notin \texttt{in}(L_2)$ then $L_1 \succ_i L_2$, otherwise, $L_1 \sim_i L_2$.*

Finally, we analyse a preference structure which is not strictly self-interested. In *aggressive preferences* an agent is interested in defeating as many arguments of other all agents' as possible, and thus does care about the labelling of arguments of others.

Definition 12 (Aggressive preferences)**.** *An agent i has* aggressive preferences *if $\forall L_1, L_2 \in \mathcal{L}$, if $|\texttt{out}(L_1) \setminus \mathcal{A}_i| \geq |\texttt{out}(L_2) \setminus \mathcal{A}_i|$ then $L_1 \succeq_i L_2$.*

4 Pareto Optimality

Welfare economics provides a formal tool for assessing outcomes in terms of how they affect the well-being of society as a whole [1]. Often these outcomes are allocations of goods or resources. In the context of argumentation, however, an outcome specifies a particular labelling. In this section, we analyse the Pareto optimality of the different argumentation outcomes. Since labellings coincide exactly with all complete extensions, in the subsequent analysis, all `in` arguments in our outcomes are conflict-free, self-defending, and contain all arguments they defend.

A key property of an outcome is whether it is *Pareto optimal*. This relies on the notion of Pareto dominance.

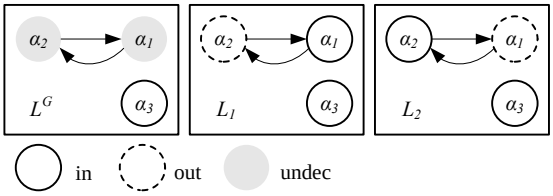
Definition 13 (Pareto Dominance)**.** *An outcome $o_1 \in \mathcal{O}$* Pareto dominates *outcome o_2 if $\forall i \in I$, $o_1 \succeq_i o_2$ and $\exists j \in I$, $o_1 \succ_j o_2$.*

An outcome is Pareto optimal if it is not Pareto dominated by any other outcome – or, equivalently, if it cannot be improved upon from one agent's perspective without making another agent worse off. Formally:

Definition 14 (Pareto Optimality)**.** *An outcome $o_1 \in \mathcal{O}$ is* Pareto optimal *(or* Pareto efficient*) if there is no other outcome $o_2 \neq o_1$ such that $\forall i \in I$, $o_2 \succeq_i o_1$ and $\exists j \in I$, $o_2 \succ_j o_1$.*

It is interesting to see that the grounded extension is *not* Pareto optimal for a population of individual acceptability maximising agents. Consider the following example.

Example 1. *Consider the graph below with three outcomes.*

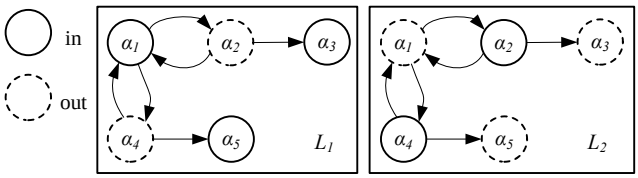


Suppose we have two agents with types $\mathcal{A}_1 = \{\alpha_1, \alpha_3\}$ and $\mathcal{A}_2 = \{\alpha_2\}$. The grounded extension is the labelling L^G, which is not Pareto optimal. Agent 1 strictly prefers L_1 and is indifferent between L^G and L_2, while agent 2 strictly prefers outcome L_2 and is indifferent between L^G and L_1.

The above observation is caused by the fact that the grounded extension is the *minimal* complete extension with respect to set inclusion. Thus, it is possible to accept more arguments without violating the fundamental requirement that the outcome is a complete extension (*i.e.* conflict-free, admissible, and includes everything it defends).

One might expect that all preferred extensions are Pareto optimal outcomes, since they are maximal with respect to set inclusion. However, as the following example demonstrates, this is not necessarily the case.

Example 2. *Consider the graph below, in which the graph has two preferred extensions.*



Suppose we have three individual acceptability maximising agents with types $\mathcal{A}_1 = \{\alpha_3, \alpha_4\}$, $\mathcal{A}_2 = \{\alpha_1\}$ and $\mathcal{A}_3 = \{\alpha_2, \alpha_5\}$. Agents $\mathcal{A}_1$ and $\mathcal{A}_3$ are indifferent between the two extensions (they get a single argument accepted in either) but agent $\mathcal{A}_2$ strictly prefers outcome L_1. Thus L_2 is not Pareto optimal.

However, it is possible to prove that every Pareto optimal outcome is a preferred extension (*i.e.* all non-preferred extensions are Pareto dominated by some preferred extension).

Theorem 1. *If agents have acceptability-maximising preferences and if an outcome is Pareto optimal then it is a preferred extension.*

Proof. Let $L \in \mathcal{L}$ be a Pareto optimal outcome. Assume that L is not a preferred extension. Since L is not a preferred extension, then there must exist a preferred extension $L^P \in \mathcal{L}$ such that $\mathtt{in}(L) \subset \mathtt{in}(L^P)$. Thus, for all i, $\mathtt{in}(L) \cap \mathcal{A}_i \subseteq \mathtt{in}(L^P) \cap \mathcal{A}_i$ and $|\mathtt{in}(L) \cap \mathcal{A}_i| \leq |\mathtt{in}(L^P) \cap \mathcal{A}_i|$ which implies that $L^P \succeq_i L$. Additionally, there exists an argument $\alpha' \in \mathcal{A}_j$ for some agent j such that $\alpha' \notin L$ and $\alpha' \in L^P$. Therefore, $|\mathtt{in}(L) \cap \mathcal{A}_j| < |\mathtt{in}(L^P) \cap \mathcal{A}_j|$ and so $L^P \succ_j L$. That is, L^P Pareto dominates L. Contradiction. □

The grounded extension turns out to be Pareto optimal for a different population of agents.

Theorem 2. *If agents have rejection-minimising preferences then the grounded extension is Pareto optimal.*

Proof. This follows from the fact that the grounded extension coincides with labellings with minimal $\texttt{out}$ labellings [3]. Thus any other outcome would have strictly more $\texttt{out}$ labels, resulting in at least one agent being made worse-off. □

It is also possible to prove the following.

Theorem 3. *If agents have rejection-minimising preferences, then for any outcome $L \in \mathcal{L}$, either L is the grounded extension, or L is Pareto dominated by the grounded extension.*

Proof. Let L^G denote the grounded extension, and let $L \in \mathcal{L}$ be any outcome. If $L = L^G$ then we are done. Assume that $L \neq L^G$. Since L^G has minimal $\texttt{out}$ among all outcomes in $\mathcal{L}$, then $\texttt{out}(L^G) \subset \texttt{out}(L)$. Thus, for each agent i, if argument $\alpha \in \mathcal{A}_i$ and $\alpha \in \texttt{out}(L^G)$ then $\alpha \in \texttt{out}(L)$. Therefore, $\texttt{out}(L^G) \cap \mathcal{A}_i \subset \texttt{out}(L) \cap \mathcal{A}_i$, and so $|\texttt{out}(L^G) \cap \mathcal{A}_i| \leq |\texttt{out}(L) \cap \mathcal{A}_i|$ which implies that $L^G \succeq_i L$. In addition, there also exists some agent j and argument α' such that $\alpha' \in \mathcal{A}_j$, $\alpha' \notin \texttt{out}(L^G)$ and $\alpha' \in \texttt{out}(L)$. Therefore, $|\texttt{out}(L^G) \cap \mathcal{A}_i| < |\texttt{out}(L) \cap \mathcal{A}_i|$ which implies that $L^G \succ_j L$. That is, L^G Pareto dominates L. □

The two previous theorems lead to a corollary.

Corollary 1. *The grounded extension characterises exactly the Pareto optimal outcome among a rejection minimising population.*

The following result relates to decisive agents.

Theorem 4. *If agents have decisive preferences, then all Pareto optimal outcomes are semi-stable extensions.*

Proof. This follows from the fact that any semi-stable extension coincides with a labelling in which $\texttt{undec}$ is minimal with respect to set inclusion [3]. The actual proof is similar in style to Theorem 1 and so due to space constraints we do not include the details. □

Note that any finite argumentation framework must have at least one semi-stable extension [4]. Moreover, when at least one stable extension exists, the semi-stable extensions are equal to the stable extensions, which themselves coincide with an empty $\texttt{undec}$ [4], which is ideal for decisive agents.

Corollary 2. *For agents with decisive preferences, if there exists a stable extension, then the stable extensions fully characterise the Pareto optimal outcomes for agents with decisive preferences.*

If a population of agents have all-or-nothing preferences then we can provide a partial characterisation of the Pareto optimal outcomes.

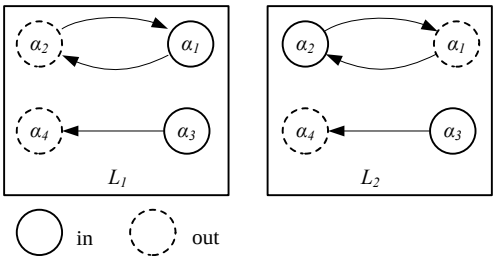
Theorem 5. *If agents have all-or-nothing preferences, then there exists a Pareto optimal preferred extension.*

Proof. We can prove this theorem by studying the possible cases. Let $\mathcal{L}$ be the set of all labellings.
Case 1: If for all $L \in \mathcal{L}$, it is the case that for all $i \in I$, $\mathcal{A}_i \not\subseteq \texttt{in}(L)$, then all agents are indifferent between all labellings, and thus all are Pareto optimal, including all preferred extensions.
Case 2: Assume there exists labelling L such that there exists an agent i with $\mathcal{A}_i \subseteq \texttt{in}(L)$ and which is Pareto optimal. If L is also a preferred extension then we are done. If L is not a preferred extension, then there must exist a preferred extension L' such that $\texttt{in}(L) \subseteq \texttt{in}(L')$. Since L was Pareto optimal, then for all agents j, it must be the case that $L \sim_j L'$ and so L' is Pareto optimal.
Case 3: Assume there exists a labelling L such that there exists an agent i with $\mathcal{A}_i \subseteq \texttt{in}(L)$ and which is not Pareto optimal. Thus, L is Pareto dominated by some labelling L^* and so there must

exist an agent j such that $\mathcal{A}_j \not\subseteq \texttt{in}(L)$ and $\mathcal{A}_i, \mathcal{A}_j \subseteq \texttt{in}(L^*)$. If L^* is not Pareto optimal then there must exist an agent k and a labelling L^{**} such that $\mathcal{A}_k \not\subseteq L^*$ and $\mathcal{A}_i, \mathcal{A}_j, \mathcal{A}_k \subseteq L^{**}$. Continue this process until the final labelling is Pareto optimal. This is guaranteed to terminate since we have a finite set of agents and labellings. Apply Case 2. □

If agents have all-or-nothing preferences, then it is possible that a preferred extension can Pareto dominate another preferred extension.

Example 3. *Consider the graph below, in which there are two preferred extensions.*



Suppose we have two agents with all-or-nothing preferences and with $\mathcal{A}_1 = \{\alpha_2, \alpha_3\}$ *and* $\mathcal{A}_2 = \{\alpha_1, \alpha_4\}$. *Outcome* L_2 *Pareto dominates outcome* L_1.

If agents have focal-argument preferences, then we can also provide a partial characterization of the Pareto optimal outcomes.

Theorem 6. *If agents have focal-argument preferences, then there exists a Pareto optimal preferred extension.*

The proof is similar to Theorem 5 and so due to space constraints we do not include the details.

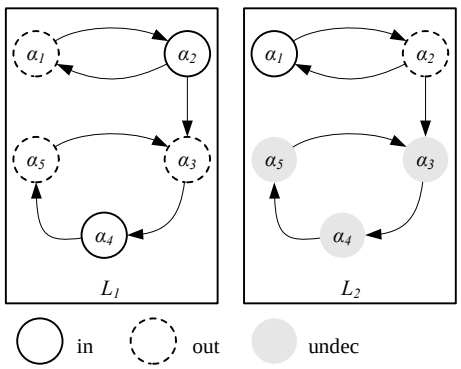
Theorem 7 says that if the population of agents have aggressive preferences, then every Pareto optimal outcome is a preferred extension.

Theorem 7. *If agents have aggressive preferences then all Pareto optimal outcomes are preferred extensions.*

Proof. Let L be a Pareto optimal outcome. Assume that L is not a preferred extension. Since L is not a preferred extension, then there must exist a preferred extension L' such that $\texttt{out}(L) \subset \texttt{out}(L')$. Thus, there must exist an agent i with $\mathcal{A}_i$ and $|\texttt{out}(L') \cap \mathcal{A}_i| > |\texttt{out}(L) \cap \mathcal{A}_i|$, and for all agents j such that $\mathcal{A}_j \in \texttt{out}(L)$, $|\texttt{out}(L') \cap \mathcal{A}_j| \geq |\texttt{out}(L) \cap \mathcal{A}_j|$ and so L' Pareto dominates L. Contradiction. □

However, not all preferred extensions are Pareto optimal, as is demonstrated in the following example.

Example 4. *Consider the graph below, in which there are two preferred extensions.*



Population Type	Pareto Optimality
Individual acceptance maximisers	Pareto optimal outcomes $\subseteq$ preferred extensions (Theorem 1)
Individual rejection minimisers	Pareto optimal outcome $=$ grounded extension (Theorem 2, 3, and Corollary 3)
Decisive	Pareto optimal outcomes $\subseteq$ semi-stable extensions (Theorem 4); if a stable extension exists, then Pareto optimal outcomes $=$ stable extensions (Corollary 2)
All-or-nothing	Some preferred extension (Theorem 5) and possibly other complete extensions
Focal argument	Some preferred extension (Theorem 6) and possibly other complete extensions
Aggressive	Pareto optimal outcomes $\subseteq$ preferred extensions (Theorem 7)

Table 2: Classical extensions & Pareto optimality

Suppose we have three agents with aggressive preferences such that $\mathcal{A}_1 = \{\alpha_2, \alpha_4\}$, $\mathcal{A}_2 = \{\alpha_1, \alpha_3\}$ *and* $\mathcal{A}_3 = \{\alpha_5\}$. *Then* $L_1 \succ_1 L_2$, $L_1 \succ_3 L_2$ *and* $L_1 \sim_2 L_2$. *That is,* L_1 *Pareto dominates* L_2.

We summarise the results from this section in Table 2. These results are important since they highlight a limitation in the definitions of extensions in classical argumentation. In some cases, Pareto optimal outcomes are fully characterised by an extension (*e.g.* grounded extension and rejection minimising agents). In other cases, however, classical extensions do not provide a full characterisation (*e.g.* for acceptance maximising agents, every Pareto optimal outcome is a preferred extension but not vice versa). In such cases, we need to explicitly refine the set of extensions in order to select the Pareto optimal outcomes (*e.g.* generate all preferred extensions, then iteratively eliminate dominated ones).

5 Restrictions on the Argumentation Framework

In Section 4 we placed no restrictions on the topological structure of the argumentation framework, nor on the structure of the argument sets of agents. In this section we impose a restriction on the argumentation framework which induces *coherency* in the framework, and then show that this provides refined characterizations of the Pareto optimal outcomes.

Definition 15 (Coherent [6])**.** *An argumentation framework,* AF, *is* coherent *if each preferred extension of* AF *is stable.*

We introduce an extended definition of defeat.

Definition 16 (Indirect defeat [6])**.** *Let* $\alpha, \beta \in \mathcal{A}$. *We say that* α indirectly defeats β *if and only if there is an odd length path from* α *to* β *in the argument graph.*

Dung introduced the notion of an argumentation framework being *limited-controversial*, which is equivalent to there being no odd-length cycles in the argumentation graph. That is, an argumentation framework is limited-controversial if no argument indirectly defeats itself. Given this restriction on the argumentation framework, the following result is obtained.

Theorem 8. *Every limited-controversial argumentation framework is coherent [6].*

Theorem 8 and Definition 15 together imply Corollary 3.

Corollary 3. *If an argumentation framework,* AF, *contains no odd-length cycles, then all of its preferred extensions are stable. That is, if* L^P *is a preferred extension of* AF *then* $\texttt{undec}(L^P) = \emptyset$.

We now introduce a restriction on the sets of arguments that agents can maintain. In particular, we assume that for each agent i, the set of arguments, $\mathcal{A}_i$, contains no arguments which indirectly

defeat each other, given the argumentation framework, AF. For example, referring to the figure in Example 2, argument α_4 indirectly defeats α_1, α_3 and α_5. Thus, we assume that no agent's argument set contains both α_4 and either α_1, α_3 or α_5. Intuitively, this property implies that each agent's arguments must be conflict-free (*i.e.* consistent), both explicitly and implicitly. Explicit consistency implies that no argument defeats another. Implicit consistency implies that other agents cannot possibly present a set of arguments that reveal an indirect defeat among one's own arguments. More concretely, exposing an indirect defeat chain can be seen as exposing a fallacy in one's arguments.

If the argument sets of the agents contain no indirect defeats with respect to the argumentation framework, then there are no odd-length cycles in the entire argument graph since, otherwise, at least one agent would have an argument that indirectly defeats itself. This allows us to provide a further characterization of the Pareto optimal outcomes for certain classes of agents' preferences.

Theorem 9. *Assume that agents have decisive preferences and that no agent has an argument set that contains indirect defeats. Then the set of stable extensions completely characterises the Pareto optimal outcomes.*

Proof. From Corollary 3 any labelling L^P that corresponds to a preferred extension must be a stable extension. From Corollary 2 if stable extensions exists then they fully characterise the set of Pareto optimal outcomes for decisive agents. □

Theorem 10. *Assume that agents have acceptability-maximising preferences, and that no agent has an argument set that contains indirect defeats. Then,*

- *there exists at least one stable extension, and*
- *every Pareto optimal outcome is a stable extension.*

Proof. Dung proved that every argumentation framework has at least one preferred extension (Corollary 12 [6]). Given the restriction on the agents' argument sets, there are no odd-length cycles and so all preferred extensions are stable (Corollary 3). By Theorem 1 if an outcome is Pareto optimal then it must be a preferred extension, and, thus, a stable extension. □

Finally, Theorems 9 and 10 allow us to characterize the Pareto optimal outcomes even when the agent population contains different preferences.

Corollary 4. *For agent populations consisting of both acceptability-maximising and decisive preferences, if agents' argument sets contain no indirect defeats with respect to the argumentation framework, then every Pareto optimal outcome is a stable extension.*

6 Further Refinement using Social Welfare

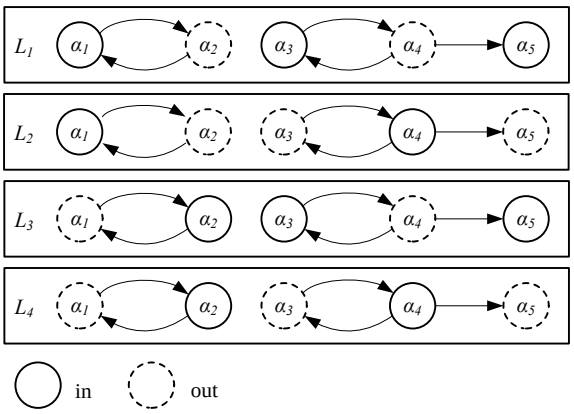
While Pareto optimality is an important way of evaluating outcomes, it does have some limitations. First, as highlighted above, there may be many Pareto optimal outcomes, and it can be unclear why one should be chosen over another. Second, sometimes Pareto optimal outcomes may be *undesirable* for some agents. For example, in a population of individual acceptability maximising agents, a preferred extension which accepts all arguments of one agent while rejecting all other arguments is Pareto optimal.

Social welfare functions provide a way of combining agents' preferences in a systematic way in which to compare different outcomes, and in particular, allow us to compare Pareto optimal extensions. We assume that an agent's preferences can be expressed by a utility function in the standard way and that it is possible to compare utility functions of the agents in a meaningful way. A social welfare function is an increasing function of individual agents' utilities and is related to the notion of Pareto optimality in that any outcome that *maximises* social welfare is also Pareto optimal.

Thus by searching for social-welfare maximising outcomes we select outcomes from among the set of Pareto optimal ones.

While there are many types of social welfare functions, two important ones are the utilitarian and egalitarian social welfare functions.[3] Example 5 illustrates how these functions can be used to compare different Pareto optimal outcomes.

Example 5. *Consider the graph below with four preferred extensions.*



Assume that there are two agents with $\mathcal{A}_1 = \{\alpha_1, \alpha_3, \alpha_5\}$ and $\mathcal{A}_2 = \{\alpha_2, \alpha_4\}$, and that these agents have acceptability maximising preferences with utility functions $u_i(L, \mathcal{A}_i) = |\mathtt{in}(L) \cap \mathcal{A}_i|$. L_1, L_3 and L_4 are all Pareto optimal. L_1 and L_3 both maximise the utilitarian social welfare, while L_3 also maximises the egalitarian social welfare function.

The above analysis shows that by taking into account welfare properties, it is possible to provide more fine grained criteria for selecting among classical extensions (or labellings) in argumentation frameworks. Such refined criteria can be seen as a sort of *welfare semantics* for argumentation.

7 Discussion and Conclusion

Until recently, argumentation-based semantics have been compared mainly on the basis of how they deal with specific benchmark problems (argument graph structures with odd-cycles *etc.*). Recently, it has been argued that argumentation semantics must be evaluated based on more general intuitive principles [2]. Our work can be seen to be a contribution in this direction. We introduced a new perspective on analysing and designing argument acceptability criteria in abstract argumentation frameworks. Acceptability criteria can now be evaluated not only based on their logically intuitive properties, but also based on their welfare properties in relation to a society of agents.

Our framework and results can be used to decide which argument evaluation rule to use given the type of agent population involved. While we formulated the problem as being *multiagent* in nature, our findings can also be extended to single-agent settings. In situations where there are several extensions, the agent can be consulted as to its preferences, in order to select the extension that the agent prefers.

The results are also of key importance to argumentation mechanism design (ArgMD) [7] where agents may argue strategically – *e.g.* possibly hiding arguments. ArgMD aims to design rules of interaction such that self-interested agents produce, in equilibrium, a particular desirable social outcome (*i.e.* the rules *implement* a particular social choice function). Understanding what social

[3]Given some outcome o, the utilitarian social welfare function returns the sum of the agents' utilities for o, while the egalitarian social welfare function returns $\min_i u_i(o, \theta_i)$.

outcomes are desirable (in this case, Pareto optimal) for different kinds of agents is an important step in the ArgMD process. Indeed, a major future research direction, opened by this paper, is the design of argumentation mechanisms that implement Pareto optimal social choice functions under different agent populations.

Acknowledgements

We thank Martin Kiefel for his contributions. He played a significant role in motivating Section 5.

References

[1] Kenneth J. Arrow, A. K. Sen, and K. Suzumura, editors. *Handbook of Social Choice and Welfare*, volume 1. Elsevier Science Publishers (North-Holland), 2002.

[2] Pietro Baroni and Massimiliano Giacomin. On principle-based evaluation of extension-based argumentation semantics. *Artificial Intelligence*, 171(10–15):675–700, 2007.

[3] Martin W. A. Caminada. On the issue of reinstatement in argumentation. In *Proceedings of the 10th European Conference on Logics in Artificial Intelligence (JELIA 2006)*, volume 4160 of *Lecture Notes in Computer Science*, pages 111–123. Springer, 2006.

[4] Martin W. A. Caminada. Semi-stable semantics. In *Proceedings of the 1st International Conference on Computational Models of Argument (COMMA)*, pages 121–130, Amsterdam, Netherlands, 2006.

[5] Sylvie Coste-Marquis, Caroline Devred, Sébastien Konieczny, Marie-Christine Lagasquie-Schiex, and Pierre Marquis. On the merging of Dung's argumentation systems. *Artificial Intelligence*, 171(10–15):730–753, 2007.

[6] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–358, 1995.

[7] Iyad Rahwan and Kate Larson. Mechanism Design for Abstract Argumentation. In *Proceedings of the 7th International Joint Conference on Autonomous Agents & Multi Agent Systems, AAMAS'2008, Estoril, Portugal*, 2008.

[8] Iyad Rahwan, Sarvapali D Ramchurn, Nicholas R. Jennings, Peter McBurney, Simon Parsons, and Liz Sonenberg. Argumentation based negotiation. *Knowledge Engineering Review*, 18(4):343–375, 2003.

[9] Jeff Rosenschein and Gilad Zlotkin. *Rules of Encounter: Designing Conventions for Automated Negotiation among Computers*. MIT Press, Cambridge MA, USA, 1994.

Kate Larson
Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
Email: klarson@cs.uwaterloo.ca

Iyad Rahwan
Faculty of Informatics, British University in Dubai
P.O. Box 502216, Dubai, UAE
(Fellow) School of Informatics, University of Edinburgh
Edinburgh, EH8 9LE, UK
Email: irahwan@acm.org

Approval-rating systems that never reward insincerity

Rob LeGrand and Ron K. Cytron

Abstract

Electronic communities are relying increasingly on the results of continuous polls to rate the effectiveness of their members and offerings. Sites such as eBay and Amazon solicit feedback about merchants and products, with prior feedback and results-to-date available to participants before they register their approval ratings. In such a setting, participants are understandably prone to exaggerate their approval or disapproval, so as to move the average rating in a favored direction.

We explore several protocols that solicit approval ratings and report a consensus outcome without rewarding insincerity. One such system is rationally optimal while still reporting an outcome based on the the usual notion of averaging. That system allows all participants to manipulate the outcome in turn. Although multiple equilibria exist for that system, they all report the same average approval rating as their outcome. We generalize our results to obtain a range of declared-strategy voting systems suitable for approval-rating polls.

1 Approval ratings and their aggregation

Approval ratings are one mechanism that communities can use to offer incentive and reward for good behavior or service. The prospect of feedback following a given interaction presumably increases the accountability of that interaction for all parties involved. Publication of approval ratings then enables appropriate consequences to follow from positive or negative experiences. In this paper we consider several forms of aggregation and we show that some methods can reward insincerity while others cannot. We next provide several examples of approval rating systems and formulate a general form of an approval rating poll.

1.1 Examples of approval rating polls

Subscribers and observers of media frequently learn of the results of *approval rating polls* that attempt to discern how strongly a participating electorate endorses a person or a position of interest. As an example, several web sites post varous forms of approval ratings for films and games. Specifically, Rotten Tomatoes [2] posts the results of two polls for each film:

- In effect, each review is turned into a 0 or 1 value, and the *Tomatometer* is the average of those values expressed as a percentage. Putative viewers might consult a film's Tomatometer value to determine whether they should see that film.
- Each critic can also rate a film's overall quality on a 1–10 scale. Rotten Tomatoes then publishes the average of all such ratings.

Finally, consider the electronic marketplace, in which participants are asked to rate the honesty and effectiveness of merchants and customers. Sites such as eBay poll their participants concerning how strongly they approve of the behavior of the marketplace members they encounter in transactions. Upon completion of a transaction, the involved parties are asked to rate each other. An aggregation of an indvidual's approval ratings is posted for public view, so that members can consider such information before engaging that individual in a transaction.

1.2 Formulation

We next define a general instance of an approval rating poll to facilitate presentation of our results.

- An electorate of n participants is polled. Based on the participants' response and the aggregation protocol at hand, the result of the poll will be published as a rational number in the interval $[0, 1]$.
- Each participant i has in mind a sincere preference rating r_i, where $0 \leq r_i \leq 1$, that can be construed as that participant's dictatorial preference. The tuple of all participants' sincere ratings is denoted by the vector $\vec{r}$. We further make the reasonable assumption that voter i's preferences are single-peaked and non-plateauing.[1]
- Finally, voter i participates in the poll by expressing a rating preference of v_i, which may or may not be the same as r_i. In fact, we are particularly interested in situations where $v_i \neq r_i$. For example, consider an eBay customer who undertakes a transaction with a highly approved merchant. If the customer becomes disgruntled with the merchant, then the customer's resulting rating of the merchant might be overly negative, precisely because of the merchant's otherwise high rating.

  The tuple of all expressed approval ratings is denoted by the vector $\vec{v}$.

This paper considers an approach that can account for, mitigate, or prevent the use of insincerity to increase a participant's effectiveness in an approval rating poll.

1.3 Aggregating approval ratings

The results of an approval rating poll are typically reported by an aggregation procedure that is disclosed *a priori*. In this section, we consider two popular aggregation schemes: average and median.

Average aggregation Here, the result of the approval rating poll is computed as the average of the participants' expressed approval ratings: $\bar{v} = \frac{\sum_{j=1}^{n} v_j}{n}$. While the Average aggregation function is sensitive to each voter's input, it has an important disadvantage: Voters can often gain by voting insincerely. For example, the 1983 film *Videodrome* has five critics' ratings on Metacritic [1]. If we assume that these critics rated the film sincerely (that each would prefer that the average rating of the film be his or her rating), we have $\vec{r} = [0.4, 0.7, 0.8, 0.8, 0.88]$. If these preferences are actually expressed sincerely in an Average aggregation context, then we have $\vec{v} = \vec{r}$ and the Average outcome is 0.716.

Consider voter 5, whose ideal outcome is $r_5 = 0.88$. That voter could achive a better outcome by *not* expressing the sincere preference $v_5 = 0.88$ and instead voting $v_5 = 1$. The resulting Average aggregation yields the outcome 0.74, which, being closer to 0.88, is preferred by voter 5 to 0.716.

Median aggregation (n odd) Another possible aggregation function computes a *median* of $\vec{v}$: $\tilde{v}$ is a value that satisfies $|\{i : \tilde{v} < v_i\}| \leq \frac{n}{2} \leq |\{i : \tilde{v} \leq v_i\}|$.[2] According to the *median*

[1] For the applications we describe, it is reasonable to assume that each voter i would prefer that the outcome be as near to the ideal r_i as possible. This single-peaked assumption makes possible the optimal strategy we describe in section 2.

[2] The above definition does not necessarily prescribe a unique outcome when n is even; we address this issue below.

voter theorem [4, 10], when n is odd, Median aggregation becomes the unique, Condorcet-compliant [13] rating system, yielding a result that is preferred by some majority of voters to every other outcome.

Unfortunately, Median aggregation can effectively ignore almost half of the voters—majority rule can mean majority tyranny. Given the tuple of votes $\vec{v} = [0, 0, 0, 1, 1]$, the 1-voters are effectively ignored when the median, 0, is chosen as the outcome. Majority tyranny could be quite undesirable for polls of this type, especially when the goal of aggregating ratings is to represent a satisfactory consensus for all voters. The Average outcome of the above tuple, 0.4, arguably provides such a much better consensus.

In contrast with Average aggregation, Median aggregation is nonmanipulable by insincere voters—at least when n is odd: a voter i can never improve the outcome from his or her point of view by voting $v_i \neq r_i$. (The treatment for an even number of voters and the proofs here and below are omitted for space; this material can be found in LeGrand [12, ch. 3].) Thus, Median aggregation does not reward insincerity for an odd number of participants.

Without losing nonmanipulability, the Median function can be generalized to give the outcome ${}^b\tilde{v}$ where $|\{i : {}^b\tilde{v} < v_i\}| \leq bn \leq |\{i : {}^b\tilde{v} \leq v_i\}|$ for any $0 \leq b \leq 1$ (in this notation, the b is intended as a parameter modifying the tilde symbol). If bn is an integer, there may be more than one $0 \leq \phi \leq 1$ that satisfies $|\{i : \phi < v_i\}| \leq bn \leq |\{i : \phi \leq v_i\}|$. In that case, define Φ as the set of all such ϕ. Then

$${}^b\tilde{v} \equiv \begin{cases} \min(\Phi) & \text{if} \quad b < \min(\Phi) \\ b & \text{if} \quad \min(\Phi) \leq b \leq \max(\Phi) \\ \max(\Phi) & \text{if} \quad \max(\Phi) < b \end{cases}$$

This order-statistic outcome equals $\max(\vec{v})$ when $b = 0$, the third quartile when $b = \frac{1}{4}$, the Median outcome when $b = \frac{1}{2}$, the first quartile when $b = \frac{3}{4}$ and $\min(\vec{v})$ when $b = 1$.

2 Rationally optimal strategy for Average aggregation

As shown in section 1, Average aggregation can reward insincerity. In this section, we develop a *rationally optimal* strategy: a computation by which a voter can achieve a result as close as possible to that voter's preferred outcome. As before, we assume an electorate in which n voters will express preferences. We begin by considering a rationally optimal ("best response") strategy from the perspective of a final, omniscient voter. We then consider the behavior of a system in which all voters use a rationally optimal strategy.

To facilitate exposition and analysis of our results, we begin by generalizing the scale on which preferences are expressed as follows. In an $[m, M]$-Average poll, voters are allowed to *express* preference ratings in the interval $[m, M]$, where $m \leq 0$ and $1 \leq M$. We continue to assume that *sincere* preference ratings are in the interval $[0, 1]$; the expanded range is therefore intended to allow voters more room to manipulate the outcome. We also assume that preferences are aggregated by computing the average of the voters' expressed preferences.

2.1 Strategy for a final, omniscient voter

Consider a $(-\infty, +\infty)$-Average poll in which voter v_n is the last voter to express an approval rating, and in which all other voters vote their sincere preference ratings: $(\forall i \neq n)\ v_i = r_i$. If voter n can see the expressed approval ratings of all voters, then the ideal outcome for voter n ($\bar{v} = r_n$) can be realized by voting $v_n = r_n n - \sum_{j \neq n} r_j$.

More generally, in an $[m, M]$-Average poll, voter n should express v_n to move the outcome

as close to r_n as possible:

$$v_n = \min\left(\max\left(r_n n - \sum_{j\neq n} r_j, m\right), M\right) \tag{1}$$

The above is the *rationally optimal strategy* for voter n in an $[m, M]$-Average approval rating poll.

As an example, consider the $[0, 1]$-Average system with sincere preferences from the *Videodrome* example above: $\vec{r} = [0.4, 0.7, 0.8, 0.8, 0.88]$. After all other voters express their sincere preferences, v_5's rationally optimal preference rating is given by

$$\begin{aligned} v_5 &= \min\left(\max\left(r_5 n - \sum_{j\neq 5} r_j, 0\right), 1\right) \\ &= \min\left(\max\left(0.88 \cdot 5 - (0.4 + 0.7 + 0.8 + 0.8), 0\right), 1\right) \;=\; 1 \end{aligned} \tag{2}$$

achieving an outcome $\bar{v}$ of 0.74. No other choice for v_5 would achieve an outcome $\bar{v}$ closer to $r_5 = 0.88$.

After voter n has voted using Equation 1, either voter n's ideal outcome r_n has been realized or voter n has moved the outcome as close to r_n as is immediately possible. Note also that $\bar{v} \in [0, 1]$ even though $v_n \in [m, M]$.

2.2 Equilibrium for n strategic voters

We have thus far allowed only voter n to use a rationally optimal strategy, requiring all other voters to express their sincere approval ratings. We now consider the properties of the more practical $[m, M]$-Average system in which each voter i uses a rationally optimal strategy to compute an expressed approval rating, based on i's sincere approval rating r_i and on the expressed votes of all other voters. When each voter i establishes v_i, other voters may wish to update their expressed approval ratings.

While there are many possible schemes that could accommodate iterative changes in expressed preferences, we examine the more general issue of reaching an equilibrium: each voter i has arrived at an expressed preference v_i such that the rationally optimal strategy recommends no change in v_i:

$$(\forall i)\; v_i = \min\left(\max\left(r_i n - \sum_{j\neq i} v_j, m\right), M\right) \tag{3}$$

So, at equilibrium, $(\forall i)\; (\bar{v} < r_i \wedge v_i = M) \vee (\bar{v} = r_i) \vee (\bar{v} > r_i \wedge v_i = m)$, and it follows that

$$(\forall i)\; \bar{v} < r_i \longrightarrow v_i = M \tag{4}$$

and

$$(\forall i)\; \bar{v} > r_i \longrightarrow v_i = m \tag{5}$$

Equation 4 says that for every i such that $\bar{v} < r_i$, $v_i = M$. So we can place a lower bound on the sum of all v_is by assuming all other v_is are at the minimum:

$$m \cdot |\{i : \bar{v} \geq r_i\}| + M \cdot |\{i : \bar{v} < r_i\}| \leq \sum_{i=1}^{n} v_i = \bar{v} n$$

Similarly, Equation 5 says that for every i such that $\bar{v} > r_i$, $v_i = m$. So we can place an upper bound on the sum of all v_is by assuming all other v_is are at the maximum:

$$\bar{v}n = \sum_{i=1}^{n} v_i \leq m \cdot |\{i : \bar{v} > r_i\}| + M \cdot |\{i : \bar{v} \leq r_i\}|$$

So we have

$$m \cdot |\{i : \bar{v} \geq r_i\}| + M \cdot |\{i : \bar{v} < r_i\}| \leq \bar{v}n \leq m \cdot |\{i : \bar{v} > r_i\}| + M \cdot |\{i : \bar{v} \leq r_i\}|$$

which implies [12, ch. 3]

$$|\{i : \bar{v} < r_i\}| \leq \frac{\bar{v} - m}{M - m} n \leq |\{i : \bar{v} \leq r_i\}|$$

Thus any average at equilibrium must satisfy the two equations

$$|\{i : \bar{v} < r_i\}| \leq \frac{\bar{v} - m}{M - m} n \tag{6}$$

and

$$\frac{\bar{v} - m}{M - m} n \leq |\{i : \bar{v} \leq r_i\}| \tag{7}$$

3 Multiple equilibria can exist

For some sincere-ratings vectors $\vec{r}$, multiple equilibria exist: there exist more than one $\vec{v}$ satisfying Equation 3. For example, if minimum vote $m = 0$, maximum vote $M = 1$ and $\vec{r} = [0.4, 0.7, 0.7, 0.8, 0.88]$ (a slight tweak to the *Videodrome* example), then any $\vec{v} = [0, v_2, v_3, 1, 1]$, where $v_2 + v_3 = 1.5$, satisfies Equation 3 and thus represents an equilibrium from which the optimal strategy would change no voter's vote.

In this case, at each possible equilibrium the outcome is $\bar{v} = 0.7$ (the ideal outcome of the two voters "conspiring" to keep it there). This is no coincidence; in general, it turns out that, even when multiple equilibria exist, the average at equilibrium is unique.

4 At most one equilibrium average rating can exist

We have seen that, given a length-n vector $\vec{r}$ of sincere ratings where $0 \leq r_i \leq 1$ for $1 \leq i \leq n$, any equilibrium $\vec{v}$ that results from every voter's using the optimal strategy will have a $\phi = \bar{v}$ that satisfies the inequalities

$$|\{i : \phi < r_i\}| \leq \frac{\phi - m}{M - m} n \tag{8}$$

and

$$\frac{\phi - m}{M - m} n \leq |\{i : \phi \leq r_i\}| \tag{9}$$

It turns out that at most one such ϕ exists for a given $\vec{r}$:

Theorem 4.1. *Given a vector $\vec{r}$ of length n where $0 \leq r_i \leq 1$ for $1 \leq i \leq n$,*

$$\begin{aligned} &|\{i : \phi_1 < r_i\}| \leq \tfrac{\phi_1 - m}{M - m} n \leq |\{i : \phi_1 \leq r_i\}| \quad \wedge \\ &|\{i : \phi_2 < r_i\}| \leq \tfrac{\phi_2 - m}{M - m} n \leq |\{i : \phi_2 \leq r_i\}| \quad \longrightarrow \quad \phi_1 = \phi_2 \end{aligned}$$

(The proof considers two symmetric cases, $\phi_1 < \phi_2$ and $\phi_2 < \phi_1$, and shows by contradiction that each is impossible.)

5 At least one equilibrium always exists

It does little good to show that all equilibria will have equal averages if an equilibrium does not always exist. Fortunately, for any set of sincere preferred outcomes $\vec{r}$, there will always be at least one equilibrium $\vec{v}$ such that no voter i would choose to change v_i according to the optimal Average strategy defined above.

We can show that a particular procedure will always find an equilibrium. We use the *Videodrome* example ($\vec{r} = [0.4, 0.7, 0.8, 0.8, 0.88]$ with $m = 0$, $M = 1$) again for demonstration. This time, let us say initial votes are assumed to be, not sincere, but *zero* (the minimum allowed vote): $\vec{v} = [0, 0, 0, 0, 0]$. Then we again allow voters to revise their votes in order, from voter 5 down to voter 1. (This particular order will prove significant.) First, voter 5 deliberates:

$$v_5 = \min\left(\max\left(r_5 n - \sum_{j\neq 5} r_j, 0\right), 1\right) = \min\left(\max\left(0.8 \cdot 5 - (0+0+0+0), 0\right), 1\right) = 1$$

and changes v_5 to 1. The voters then in turn reason similarly and change v_4 to 1, v_3 to 1, v_2 to 0.5 and v_1 to 0. The resulting vote vector, $\vec{v} = [0, 0.5, 1, 1, 1]$, is indeed the same equilibrium found above in section 2.2, this time going through the voters only once.

This procedure inspires the following straightforward algorithm, which takes a $\vec{r}$ as input and outputs an equilibrium $\vec{v}$, assigning to each v_i exactly once. It orders the voters by decreasing r_i values, then uses the optimal strategy for each voter i in order, implicitly making the assumption that $v_j = m$ for $j > i$.

Algorithm 5.1. *FindEquilibrium*$(\vec{r}, m, M)$:
sort $\vec{r}$ so that $(\forall i \leq j)\ r_i \geq r_j$
for $i = 1$ to n do
$v_i \leftarrow \min\left(\max\left(r_i n - \sum_{k<i} v_k - (n-i)m, m\right), M\right)$
return $\vec{v}$

Note that the algorithm assigns a value between m and M, inclusive, to each v_i exactly once, and that the assignment to v_i does not depend on the values of v_j where $j > i$. Therefore, after Algorithm 5.1 completes, it must be true that

$$(\forall i)\ v_i = \min\left(\max\left(r_i n - \sum_{k<i} v_k - (n-i)m, m\right), M\right)$$

but this is not quite enough to see that the resulting $\vec{v}$ is an equilibrium. To see that, we must show that an intermediate voter would not change his or her vote even after later voters have voted:

Theorem 5.2. *For any $\vec{r}$, where $0 \leq r_i \leq 1$ for $1 \leq i \leq n$, the vote vector $\vec{v}$ returned by Algorithm 5.1 satisfies*

$$(\forall i)\ v_i = \min\left(\max\left(r_i n - \sum_{k\neq i} v_k, m\right), M\right)$$

(The proof essentially shows that, because of the way that Algorithm 5.1 orders the voters, a certain kind of "partial" equilibrium is satisfied after each step of the algorithm, which implies that an equilibrium is found after the last step.)

So an equilibrium $\vec{v}$ must always exist for any input $\vec{r}$ and any $m \leq 0$ and $M \geq 1$.

We now know that, given some sincere-preference vector $\vec{r}$,

- at most one value ϕ satisfies Equations 8 and 9 (Theorem 4.1),
- any equilibrium $\vec{v}$ has average vote $\bar{v}$ satisfying Equations 6 and 7 (section 2.2), and
- at least one equilibrium $\vec{v}$ must exist (Theorem 5.2)

and so we can conclude that any ϕ that satisfies Equations 8 and 9 must equal the average vote $\bar{v}$ at all possible equilibria $\vec{v}$.

6 Average-Approval-Rating DSV

We have seen that Algorithm 5.1, *FindEquilibrium*, always finds an equilibrium for any sincere-preference vector $\vec{r}$. We also know that any equilibrium $\vec{v}$ will have the same average $\bar{v}$ (and that $0 \leq \bar{v} \leq 1$). It follows that the average at equilibrium is unique and can be defined as a function:

Algorithm 6.1. *AverageAtEquilibrium*$(\vec{r}, m, M)$:
$\vec{v} \leftarrow$ FindEquilibrium$(\vec{r}, m, M)$
return $\bar{v} = \frac{\sum_{i=1}^{n} v_i}{n}$

Even when $m < 0$ and/or $M > 1$, *AverageAtEquilibrium* will return an outcome between 0 and 1. In fact, the outcome returned will be within the range defined by the input vector of cardinal preferences:

Theorem 6.2. $(\forall m \leq 0, M \geq 1)\ \min(\vec{r}) \leq \text{AverageAtEquilibrium}(\vec{r}, m, M) \leq \max(\vec{r})$.

6.1 Declared-Strategy Voting

In 1996, Lorrie Cranor and Ron K. Cytron [9] described a hypothetical voting system they called Declared-Strategy Voting (DSV). DSV can be seen as a meta-voting system, in that it uses voters' expressed preferences among alternatives to vote rationally in their stead in repeated simulated elections. The repeated simulated elections are run according to the rules of some underlying voting protocol, which can be any protocol that accepts any kind of ballots and uses them to choose one outcome. Cranor [8] explored using DSV with plurality, but DSV, as a meta-voting system, could conceivably work with any voting protocol for which a rationally optimal strategy can be described, such as Average aggregation.

6.2 A new class of rating systems

The Average and Median protocols necessarily take a vote vector $\vec{v}$ as input—voters' sincere preference information cannot be directly and reliably elicited, so $\vec{r}$ is not generally available. If the Average system is used and voters are rationally strategic (and are allowed to keep changing their votes until all decide to stop), the outcome can reasonably be expected to equal AverageAtEquilibrium$(\vec{r}, 0, 1)$. But instead of using Average on the vote vector $\vec{v}$ and relying on the voters to use optimally rational strategy when deciding on their votes v_i, AverageAtEquilibrium$(\vec{v}, 0, 1)$ can be calculated and taken as the outcome, implicitly and effectively using the DSV framework with Average as the underlying voting protocol. In fact, we are not limited to AverageAtEquilibrium$(\vec{v}, 0, 1)$; AverageAtEquilibrium$(\vec{v}, m, M)$ lies between 0 and 1 for any $m \leq 0$ and $M \geq 1$ and so can serve as a rating system as well.

For illustration, we reuse the *Videodrome* example and assume sincere voters: $\vec{v} = [0.4, 0.7, 0.8, 0.8, 0.88]$. Suppose we want to take as the outcome of this election not the

average vote $\bar{v}$ or the median vote $\tilde{v}$ but AverageAtEquilibrium$(\vec{v}, 0, 1)$. First we calculate FindEquilibrium$(\vec{v}, 0, 1)$, which we have seen in section 2.2 to be

$$\vec{w} = \text{FindEquilibrium}(\vec{v}, 0, 1) = [0, 0.5, 1, 1, 1]$$

Then we see that

$$\bar{w} = \frac{\sum_{i=1}^{5} w_i}{5} = \frac{0 + 0.5 + 1 + 1 + 1}{5} = 0.7$$

giving the outcome as 0.7, which equals neither the Average outcome ($\bar{v} = 0.716$) nor the Median outcome ($\tilde{v} = 0.8$).

Alternatively, we can let $m = -99$ and $M = 100$. Then the equilibrium we find turns out to be

$$\vec{w} = \text{FindEquilibrium}(\vec{v}, -99, 100) = [-99, -99, 2, 100, 100]$$

And then

$$\bar{w} = \frac{\sum_{i=1}^{5} w_i}{5} = \frac{-99 + (-99) + 2 + 100 + 100}{5} = 0.8$$

This time the effective power to determine the outcome fell to voter 3 rather than voter 2, giving the Median outcome of 0.8. (We will see that if $m + M = 1$ and $M - m$ is allowed to become large enough, the resultant outcome will equal the Median outcome.)

It turns out that in neither of these cases will any voter be able to gain from voting insincerely. This is no coincidence.

This Average-Approval-Rating (AAR) DSV system has three intuitively desirable properties: a kind of monotonicity (Theorem 6.3), immunity to Average-like strategy (Theorem 6.4) and a general nonmanipulability (Theorem 6.5). The first two will imply the third.

6.3 Monotonicity of AAR DSV

First, the monotonicity property: When some input votes are increased and none is decreased, the outcome never decreases.

Theorem 6.3. *If* $\vec{v} = [v_1, v_2, \ldots v_n]$ *and* $\vec{v}' = [v'_1, v'_2, \ldots v'_n]$ *where* $(\forall i)\ v_i \leq v'_i$, *then* AverageAtEquilibrium$(\vec{v}, m, M) \leq$ AverageAtEquilibrium$(\vec{v}', m, M)$.

(The proof is by contradiction.)

6.4 AAR DSV is immune to Average-style strategy

Another desirable property of AAR DSV is that its outcome is unaffected by voters' using Average-style strategy, trying to move the outcome in the desired direction by moving their votes in that direction.

Theorem 6.4. *If* $\vec{v} = [v_1, v_2, \ldots v_n]$ *and* $\vec{v}' = [v'_1, v'_2, \ldots v'_n]$ *where, for all* $1 \leq i \leq n$,

- $v'_i \leq v_i$ *if* AverageAtEquilibrium$(\vec{v}, m, M) > v_i$
- $v'_i = v_i$ *if* AverageAtEquilibrium$(\vec{v}, m, M) = v_i$
- $v'_i \geq v_i$ *if* AverageAtEquilibrium$(\vec{v}, m, M) < v_i$

then AverageAtEquilibrium$(\vec{v}', m, M) =$ AverageAtEquilibrium$(\vec{v}, m, M)$.

(The proof relies on Theorem 4.1.)

6.5 AAR DSV never rewards insincerity

For any voting system, it is desirable to show that a voter can never gain a better outcome by voting insincerely than by voting sincerely, however sincerity is defined. It turns out that, when AverageAtEquilibrium$(\vec{v}, m, M)$ is selected as the outcome, no voter i can gain an outcome closer to the ideal r_i by voting $v_i \neq r_i$ instead of $v_i = r_i$, guaranteeing a strong nonmanipulability property to AAR DSV:

Theorem 6.5. *If* $\vec{v} = [v_1, v_2, \ldots v_n]$ *where* $v_1 = r_1$ *and* $\vec{v}' = [v'_1, v'_2, \ldots v'_n]$ *where* $v'_1 \neq r_1$ *and* $(\forall i > 1)\ v'_i = v_i$, *then* $|\text{AverageAtEquilibrium}(\vec{v}, m, M) - r_1| \leq |\text{AverageAtEquilibrium}(\vec{v}', m, M) - r_1|$.

(The proof consists of four cases and relies on Theorems 6.3 and 6.4.)

7 Evaluation of AAR DSV systems

To simplify the evaluation of AAR DSV systems, we re-parameterize them by defining

$$\Phi_{a,b}(\vec{v}) \equiv \lim_{x \to a^+} \text{AverageAtEquilibrium}\left(\vec{v}, b - \frac{b}{x}, b + \frac{1-b}{x}\right)$$

(The limit is needed for the $a = 0$ case; as a approaches 0, $\Phi_{a,b}(\vec{v})$ approaches the ${}^b\tilde{v}$ outcome defined in section 1.3.)

Any system that uses the outcome function $\Phi_{a,b}(\vec{v})$ where $0 \leq a \leq 1$ and $0 \leq b \leq 1$ has the property that no voter can gain by voting insincerely. But it does not follow that any values of a and b give equally desirable outcomes.

One approach to evaluating this continuous range of nonmanipulable systems is to take the Average system as a benchmark and determine which $\Phi_{a,b}$ function comes nearest, on average, to giving the Average outcome. Given a vote vector $\vec{v}$, we can calculate the Average outcome $\bar{v}$ and the outcome $\Phi_{a,b}(\vec{v})$ for many a, b combinations. For any particular a and b, we can calculate the squared error from $\bar{v}$: $\text{SE}_{a,b}(\vec{v}) = (\Phi_{a,b}(\vec{v}) - \bar{v})^2$. If $\mathbf{V} = \{\vec{v}_1, \vec{v}_2, \vec{v}_3 \ldots \vec{v}_N\}$ is a vector of N vote vectors, then we can find the root-mean-squared error from Average, weighted by the number of ratings in each vote vector $\vec{v}_i$:

$$\text{RMSE}_{a,b}(\mathbf{V}) = \sqrt{\frac{\sum_{i=1}^{N} |\vec{v}_i| \cdot \text{SE}_{a,b}(\vec{v}_i)}{\sum_{i=1}^{N} |\vec{v}_i|}}$$

Given some "training" vector $\mathbf{V}$ of vote vectors, we would like to choose a and b to minimize $\text{RMSE}_{a,b}(\mathbf{V})$.

This approach requires a concrete source of vote-vector data or a distribution for generating such. The website Metacritic [1] offers ideal data for our purposes: Reviews for over 4000 films are summarized into ratings between 0 and 100. For example, one film[3] has the seven ratings 70, 70, 80, 80, 88, 88 and 100, which are easily converted into the vote vector $\vec{v} = [0.7, 0.7, 0.8, 0.8, 0.88, 0.88, 1]$. Converting all films on Metacritic the same way gives us a large vector $\mathbf{V}$ of vote vectors.[4]

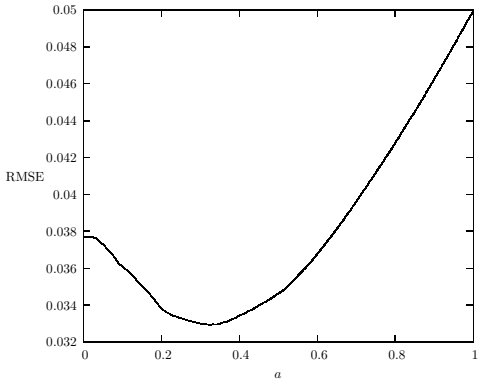
Since there are two parameters, a and b, it is somewhat impractical to try all combinations. But it may be desired to fix $b = 0.5$ to ensure a kind of symmetry: If $(\forall i)\ v'_i = 1 - v_i$,

[3] The 1978 film *Animal House*.

[4] We use the data for the 4581 films mined from Metacritic on Thursday, 3 April 2008, that had at least three critics rate them. Note that we are implicitly assuming that the rating data are sincere; unfortunately, we know of no large data set gathered using a nonmanipulable rating protocol such as ours, so we must hope that most critics are more interested in maintaining their professional reputations than in optimizing a film's Metacritic rating.

then $(\forall a)\ \Phi_{a,0.5}(\vec{v}') = 1 - \Phi_{a,0.5}(\vec{v})$, so electorates that prefer low and high outcomes are treated symmetrically. Fixing $b = 0.5$ and trying all 10001 evenly spaced values of a, we find that $a = 0.3240$ (Figure 1) gives the minimum RMSE for the Metacritic data.

Figure 1: RMSE, varying a and fixing $b = 0.5000$



Having fixed $b = 0.5$ and found the value of a that minimizes RMSE (0.3240), we can now fix $a = 0.3240$ and find the value of b that minimizes RMSE, then fix b again accordingly and continue in a hill-climbing fashion until we find a stable minimum. In practice, the procedure is guaranteed to halt because the RMSE decreases at each step for which either a or b changes.

Using this procedure on the Metacritic data and testing 10001 evenly spaced values of a or b at each step, whether we start with $a \in \{0, 0.25, 0.5, 0.75\}$ or with $b \in \{0, 0.25, 0.5, 0.75\}$,[5] we find a local RMSE minimum (approximately 0.03242) at $a = 0.3647$, $b = 0.4820$; such a system is equivalent to running an Average election with rationally optimal voters and allowing votes between $m \approx -0.8396$ and $M \approx 1.9023$.

Other preference domains may have very different properties and thus different ideal values for a and b.

8 Related and future work

In this paper we have applied the DSV framework of Cranor and Cytron [9] to create a large class of nonmanipulable rating systems, assuming only that each voter has a single-peaked preference function over the bounded, one-dimensional outcome space. The single-peaked assumption allows us to avoid the negative implications of the Gibbard-Satterthwaite theorem [11, 15], making it possible to find nonmanipulable protocols that have no dictator.

Most relevant to our work is Moulin's [14] result. He characterized the set of all non-manipulable protocols that resolve a vector of real inputs into one real outcome, showing that any such protocol is equivalent to adding some fixed set of $n - 1$ points to the n input points and taking as the outcome the median of the combined set. It turns out that our AverageAtEquilibrium$(\vec{v}, m, M)$ is equivalent to adding the (evenly spaced) points

$$m + \frac{1}{n}(M - m), m + \frac{2}{n}(M - m), \ldots m + \frac{i}{n}(M - m), \ldots m + \frac{n-1}{n}(M - m)$$

[5]Note that when $a = 1$, the outcome is simply AverageAtEquilibrium$(\vec{v}, 0, 1)$ and does not depend on the the b parameter, so different values of b cannot be compared. When b is set to 1, RMSE turns out to be minimized at $a = 1$.

to the input points (which fall between 0 and 1) and taking the median of that set as the outcome, so our set of AAR DSV systems is indeed a subset of Moulin's set of nonmanipulable protocols.

In future work we plan to explore higher-dimensional outcome spaces. The Median system can be perhaps most naturally generalized to $d > 1$ dimensions by finding the point t that minimizes $\sum_{i=1}^{n} \text{dist}(t, v_i)$, where $\text{dist}(t, v_i) = \left(\sum_{j=1}^{d} (t_j - v_{ij})^2\right)^{1/2}$, the Euclidean distance between t and v_i. t is known as the Fermat–Weber point [18, 7]. When $d > 1$, unlike in the one-dimensional case, it usually has a single optimum point even when n is even (the only exception is an even number of collinear points). Unfortunately, there is no computationally feasible exact algorithm to calculate the Fermat–Weber point in general [3], but numerical approximation is quite easy [17, 6].

The Fermat–Weber point does not change when a point v_i is moved farther away from t in the direction of the vector from t to v_i [16], so, in a sense, direction matters but not distance. Because of this property, a naïve Average-style strategy for manipulating this Fermat–Weber system fails, and any successful manipulation would have to move a sincere vote in some other direction. Unfortunately, an insincere voter can indeed manipulate the Fermat–Weber point to move closer to his or her ideal outcome [12, ch. 3]. In fact, Zhou [19] showed that no protocol with effective outcome space of dimension $d > 1$ is generally nonmanipulable when voters can have any single-peaked (concave) preference function.

The Average system is easily generalized to higher-dimension hypercubes by taking the average of each coordinate, effectively calculating the centroid, the center of mass given a set of unit masses. This generalization is equivalent to finding the point t that minimizes $\sum_{i=1}^{n} \text{dist}(t, v_i)^2$. The resulting system is equivalent to conducting d separate and independent Average elections, and the results above for strategic behavior under the one-dimensional Average system apply to the "election" for each coordinate. In particular, if each voter has *separable* preferences [5] (preferences in one dimension are independent of preferences in all other dimensions), conducting a d-dimensional AAR DSV election is equivalent to conducting d parallel one-dimensional AAR DSV elections, and so gives a nonmanipulable system. Such a preference-function space is not *abundant* by Zhou's definition.

The one-dimensional space between 0 and 1 can be generalized in other ways than into hypercubes. For example, the outcome space could be the d-dimensional simplex (for example, $\{(x, y, z) \in \mathbb{R}^3 : x + y + z = 1\}$), which could describe the division of a limited resource among several uses (such as a committee allocating a fixed sum among budget items). Unfortunately, even when all voters' preferences are separable, AAR DSV systems may be manipulable—in a sense, dimensions are interdependent for the outcome space itself. It may be, however, that no voter can move the outcome to one which is closer to ideal on one dimension without moving it further on some other dimension. We plan to investigate this "dominance"-nonmanipulability.

We would like to thank Steven Brams and the anonymous reviewers for their valuable comments and suggestions on this work.

References

[1] Metacritic. `http://www.metacritic.com/`.

[2] Rotten Tomatoes. `http://www.rottentomatoes.com/`.

[3] Chandrajit L. Bajaj. The algebraic degree of geometric optimization problems. *Discrete & Computational Geometry*, 3:177–191, 1988.

[4] Duncan Black. On the rationale of group decision-making. *Journal of Political Economy*, 56(1):23–34, February 1948.

[5] Kim C. Border and J. S. Jordan. Straightforward elections, unanimity and phantom voters. *The Review of Economic Studies*, 50(1):153–170, January 1983.

[6] Prosenjit Bose, Anil Maheshwari, and Pat Morin. Fast approximations for sums of distances, clustering and the Fermat-Weber problem. *Computational Geometry: Theory and Applications*, 24(3):135–146, April 2003.

[7] Jack Brimberg. The Fermat-Weber problem revisited. *Mathematical Programming*, 71:71–76, 1995.

[8] Lorrie Faith Cranor. *Declared-Strategy Voting: An Instrument for Group Decision-Making*. PhD thesis, Department of Computer Science, Washington University, St. Louis, Missouri, December 1996.

[9] Lorrie Faith Cranor and Ron K. Cytron. Towards an information-neutral voting scheme that does not leave too much to chance. In *Midwest Political Science Association Annual Meeting*, pages 1–15, April 1996.

[10] Anthony Downs. *An Economic Theory of Democracy*. Harper, New York, New York, 1957.

[11] Allan Gibbard. Manipulation of voting schemes: A general result. *Econometrica*, 41(3):587–601, May 1973.

[12] Rob LeGrand. *Computational Aspects of Approval Voting and Declared-Strategy Voting*. PhD thesis, Department of Computer Science and Engineering, Washington University, St. Louis, Missouri, April 2008.

[13] Samuel Merrill III. *Making Multicandidate Elections More Democratic*. Princeton University Press, Princeton, New Jersey, 1988.

[14] Hervé Moulin. On strategy-proofness and single peakedness. *Public Choice*, 35(4):437–455, January 1980.

[15] Mark Allen Satterthwaite. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10(2):187–217, April 1975.

[16] Christopher G. Small. A survey of multidimensional medians. *International Statistical Review*, 58(3):263–277, December 1990.

[17] Y. Vardi and Cun-Hui Zhang. A modified Weiszfeld algorithm for the Fermat-Weber location problem. *Mathematical Programming*, 90(3):559–566, 2001.

[18] Alfred Weber. *Alfred Weber's Theory of the Location of Industries*. University of Chicago Press, Chicago, Illinois, 1929. Translated from the German by C. J. Freidrich.

[19] Lin Zhou. Impossibility of strategy-proof mechanisms in economies with pure public goods. *The Review of Economic Studies*, 58(1):107–119, January 1991.

Rob LeGrand Ron K. Cytron
Department of Computer Science & Engineering
Washington University
One Brookings Drive, Campus Box 1045
St. Louis, Missouri 63130-4899
Email: {legrand,cytron}@cse.wustl.edu

Dodgson's Rule Approximations and Absurdity

John C. M^{c}Cabe-Dansted[1]

Abstract

With the Dodgson rule, cloning the electorate can change the winner, which Young (1977) considers an "absurdity". Removing this absurdity results in a new rule (Fishburn, 1977) for which we can compute the winner in polynomial time (Rothe et al., 2003), unlike the traditional Dodgson rule. We call this rule DC and introduce two new related rules (DR and D&). Dodgson did not explicitly propose the "Dodgson rule" (Tideman, 1987); we argue that DC and DR are better realizations of the principle behind the Dodgson rule than the traditional Dodgson rule. These rules, especially D&, are also effective approximations to the traditional Dodgson's rule. We show that, unlike the rules we have considered previously, the DC, DR and D& scores differ from the Dodgson score by no more than a fixed amount given a fixed number of alternatives, and thus these new rules converge to Dodgson under any reasonable assumption on voter behaviour, including the Impartial Anonymous Culture assumption.

1 Introduction

Finding the Dodgson winner to an election can be very difficult, Bartholdi et al. (1989) proved that determining whether an alternative is the Dodgson winner is an NP-hard problem. Later Hemaspaandra et al. (1997) refined this result by showing that the Dodgson winner problem was complete for parallel access to NP, and hence not in NP unless the polynomial hierarchy collapses. This result was of interest to computer science as the previously known problems in this complexity class were obscure by comparison.

For real world elections we do not want intractable problems. Tideman (1987) proposed a simple rule to approximate the Dodgson rule. The impartial culture assumption states that all votes are independent and equally likely. Under this assumption, it has been proven that the probability that Tideman's rule picks the Dodgson winner converges to one as the number of voters goes to infinity (M^{c}Cabe-Dansted et al., 2007). Our paper also showed that this growth was not exponentially fast. Two rules have been independently proposed for which this convergence is exponentially fast, our Dodgson Quick (DQ) rule and the `GreedyWinner` algorithm proposed by Homan and Hemaspaandra (2005). It is usually easy to verify that the DQ winner and GreedyWinner are the same as the Dodgson winner, a property that Homan and Hemaspaandra formalise as being "frequently self-knowingly correct". However the proofs of convergence depended heavily on the unrealistic impartial culture assumption. We shall show that under the Impartial Anonymous Culture (IAC) these rules do not converge.

The importance of ensuring that statistical results hold on reasonable assumptions on voter behaviour is considered by Procaccia and Rosenschein (2007). They define "deterministic heuristic polynomial time algorithm" in terms of both the problem to be solved and a probability distribution over inputs. However they do not consider the issue of whether a heuristic is self-knowingly correct. Thus we extend the concept of an algorithm being "frequently self-knowingly correct" to allow particular probability distributions to be specified.

[1]The author would like to thank Arkadii Slinko for his many valuable suggestions and references.

Procaccia and Rosenschein (2007) also propose "Junta" distributions; these distributions are intended to produce problems that are harder than would be produced under reasonable assumptions on voter behaviour. Hence if it is easy to solve a problem when input is generated according to a Junta distribution it is safe to assume that it will be easy under any reasonable assumption of voter behaviour. We will not use Junta distributions, but instead simply use that fact that even "neck-and-neck" national elections are won by thousands of votes. We will also show that the rules we have considered previously, Tideman, Dodgson Quick etc., do not converge to Dodgson's rule under IAC.

However, the reason that the Dodgson rule is so hard to compute is because cloning the electorate can change the winner. That is, if we replace each vote with two (or more) identical votes, this may change the winner. When discussing majority voting Young (1977) described this property as an "absurdity". Young suggested that such absurdities be fixed in majority voting by allowing fractions of a vote to be deleted. Fishburn (1977) proposed a similar modification to the traditional Dodgson rule, which we call Dodgson Clone. The Dodgson Clone scores can be computed by relaxing the integer constraints on the Integer Linear Program that Bartholdi et al. (1989) proposed to calculate the Dodgson score; normal (rational) Linear Programs can be solved in polynomial time, we can compute the Dodgson Clone score in polynomial time (Rothe et al., 2003).

The Dodgson Clone rule is also an effective approximation to the Dodgson rule. In computer science, an approximation typically refers to an algorithm that selects a value that is always accurate to within some error. This form of approximation is not meaningful when selecting a winner, although these rules can be used to approximate the frequency that Dodgson winner has some property. For example, Shah (2003) used Tideman's rule to approximate the frequency that the Dodgson winner matched the winners according to other rules, and so such Tideman, DQ etc. can be considered approximations of Dodgson's rule in a loose sense. However we can approximate the Dodgson score. We will show that for a fixed number of alternatives, the Dodgson Clone score approximates the Dodgson score to within a constant error. As it is implausible that the margin by which the winner wins the election will not grow with the size of the electorate, the Dodgson winner will converge to the Dodgson Clone winner under any reasonable assumption of voter behaviour. In particular we will show that they will converge under the Impartial Anonymous Culture assumption.

We propose two closer approximations Dodgson Relaxed (DR) and Dodgson Relaxed and Rounded (D&). These approximations, like the traditional Dodgson rule, are not resistant to cloning the electorate. This allows them to be closer to the Dodgson rule than Dodgson Clone, both rules converge to the Dodgson rule exponentially quickly under the Impartial Culture assumption. The DR rule is superior to the Dodgson rule in the sense that it can split ties in favour of alternatives that are fractionally better. The D& scores are rounded up, so the D& rule does not have this advantage. However it is exceptionally close to Dodgson. In 43 million elections randomly generated according to various assumptions on voter behaviour, the D& winner differed from the Dodgson winner in only one election.

The approximation proposed by Procaccia et al. (2007) is similar to these approximations in the sense that it involves a relaxation of the integer constraints. However their approximation is randomised, and thus quite different from our deterministic approximations. Using a randomised approximation as a voting rule would be unusual, and they do not discuss the merits of such a rule. Thus the focus of their paper is quite different, as we present rules that we argue are superior to the traditional formalisation of the Dodgson rule. Additionally, they do not discuss the issue of frequently self-knowing correctness.

Another approach to computing the Dodgson score has been to limit some parameter. Bartholdi et al. (1989) showed that computing the Dodgson scores and winner is polynomial when either the number of voters or alternatives is limited. It was shown that computing these from a voting situation is logarithmic with respect to the number of voters when the

number of alternatives is fixed (M$^{\text{c}}$Cabe-Dansted, 2006), and hence Dodgson winner is Fixed Parameter Tractable (FPT) with number of alternatives as the fixed parameter. It is now also known that the Dodgson winner is FPT when the Dodgson score is taken as the fixed parameter (Betzleri et al., 2008).

Thus we will define the Dodgson based rules in terms of Condorcet-tie winners, rather than Condorcet winners. As we will discuss briefly, this does not affect convergence.

2 Preliminaries

In our results we use the term agent in place of voter and alternative in place of candidate, as not all elections are humans voting other humans into office. For example, in direct democracy, the citizens vote for laws rather than candidates.

We assume that agents' preferences are transitive, i.e. if they prefer a to b and prefer b to c they also prefer a to c. We also assume that agents' preferences are strict, if a and b are distinct they either prefer a to b or b to a. Thus we may consider each agent's preferences to be a ranking of each alternative from best to worst.

Let $\mathcal{A}$ and $\mathcal{N}$ be two finite sets of cardinality m and n respectively. The elements of $\mathcal{A}$ will be called alternatives, the elements of $\mathcal{N}$ agents. We represent a **vote** by a linear order of the m alternatives. We define a **profile** to be an array of n votes, one for each agent. Let $\mathcal{P} = (P_1, P_2, \ldots, P_n)$ be our profile. If a linear order $P_i \in \mathcal{L}(A)$ represents the preferences of the i^{th} agent, then by aP_ib, where $a, b \in \mathcal{A}$, we denote that this agent prefers a to b.

A multi-set of linear orders of $\mathcal{A}$ is called a **voting situation**. A voting situation specifies which linear orders were submitted and how many times they were submitted but not who submitted them.

The Impartial Culture (IC) assumption is that each profile is equally likely. The Impartial Anonymous Culture (IAC) assumption is that each voting situation is equally likely. To understand the difference, consider a two alternative election with billions of agents; under IC it is almost certain that each alternative will get 50% ($\pm$0.5%) of the vote; under IAC, 50.0% is no more likely than any other value.

Let $\mathcal{P} = (P_1, P_2, \ldots, P_n)$ be our profile. We define n_{xy} to be the number of linear orders in $\mathcal{P}$ that rank x above y, i.e. $n_{xy} \equiv \#\{i \mid xP_iy\}$.

Definition 2.1. *The **advantage** of a over b is defined as follows:*

$$adv(a, b) = \max(0, n_{ab} - n_{ba})$$

*A **Condorcet winner** is an alternative a for which $adv(a, b) > 0$ for all other alternatives b. We define a **Condorcet-tie winner**, to be an alternative a such $adv(b, a) = 0$ for all other alternatives a. A Condorcet winner or Condorcet-tie winner does not always exist.*

It is traditional to define the Dodgson score of an alternative as the terms of the minimum number of swaps of neighbouring alternatives required to make that alternative *defeat* all others in pairwise elections, i.e. make the alternative a Condorcet winner. When not requiring solutions to be integer this becomes undefined, as if we defeat an alternative by $\epsilon > 0$ then there exists a better solution where we defeat the alternative by only $\epsilon/2$.

For this reason, when defining the Dodgson scores we only require that the alternative defeat *or tie* other alternatives, i.e. make the alternative a Condorcet-tie winner. For better consistency with the more traditional Dodgson rule we could define the Condorcet winner as an alternative a for which $\text{adv}(a, b) \geq 1$. However this would mean that the Dodgson Clone rule would not be resistant to cloning of the electorate.

This difference in definition does not affect convergence. Our proof of convergence relies only on fact that Dodgson, D&, DR and DC scores differ by at most a fixed amount

($\mathcal{O}(m!)$) when the number of alternatives is fixed. To convert a Condorcet-tie winner c into a Condorcet winner c we need to swap c over at most $(m-1)$ alternatives, each requiring at most $(m-1)$ swaps of neighbouring alternatives. Hence the difference between the score according to these different definitions of Dodgson is at most $(m-1)^2$.

We will now define a number of rules in terms of scores. The winner of each rule below is the alternative with the lowest score.

The **Dodgson score** (Dodgson 1876, see e.g. Black 1958; Tideman 1987), which we denote as $\mathrm{Sc}_{\mathbf{D}}(a)$, of an alternative a is defined as the minimum number of swaps of neighboring alternatives required to make a a Condorcet-tie winner. We call the alternative(s) with the lowest Dodgson score(s) the **Dodgson winner**(s). (Bartholdi et al., 1989)

The **Tideman score** $\mathrm{Sc}_{\mathbf{T}}(a)$ of an alternative a is:

$$\mathrm{Sc}_{\mathbf{Q}}(a) = \sum_{b \neq a} \mathrm{adv}(b, a).$$

The **Dodgson Quick (DQ) score** $\mathrm{Sc}_{\mathbf{Q}}(a)$, of an alternative a is

$$\mathrm{Sc}_{\mathbf{Q}}(a) = \sum_{b \neq a} F(b, a), \text{ where } F(b, a) = \left\lceil \frac{\mathrm{adv}(b, a)}{2} \right\rceil.$$

Although the definitions of $\mathrm{Sc}_{\mathbf{Q}}$ and $\mathrm{Sc}_{\mathbf{T}}$ are very similar, the Dodgson Quick rule converges exponentially fast to Dodgson's rule under the Impartial Culture assumption, where as Tideman's rule does not (M$^{\mathrm{c}}$Cabe-Dansted et al., 2007). This is because Dodgson and DQ are more sensitive to a large number of alternatives defeating a by a small odd margin (e.g. 1) than Tideman is.

We define the k-Dodgson score $\mathrm{Sc}^k_{\mathbf{D}}(d)$ of a as being the Dodgson score of a in a profile where each agent has been replaced with k clones, divided by k. That is, where $\mathcal{P}$ is our fixed profile, $\mathcal{P}^k$ is the profile with each agent replaced with k clones, and $\mathrm{Sc}_{\mathbf{D}}[\mathcal{P}](a)$ is the Dodgson score of a in the profile $\mathcal{P}$, then

$$\mathrm{Sc}^k_{\mathbf{D}}(a) = \mathrm{Sc}^k_{\mathbf{D}}[\mathcal{P}](a) = \frac{\mathrm{Sc}_{\mathbf{D}}[\mathcal{P}^k](a)}{k}$$

We define the **Dodgson Clone (DC) score** $\mathrm{Sc}_{\mathbf{C}}(a)$ of an alternative a as $\min_k \mathrm{Sc}^k_{\mathbf{D}}(a)$. The DC score can be equivalently defined by modifying the Dodgson rule to allow votes to be split into rational fractions and allowing swaps to be made on those fractions of a vote. Note that like the Tideman approximation, the DC score is less sensitive than Dodgson to a large number of alternatives defeating a by a margin of 1, so the DC score is unlikely to converge to Dodgson as quickly as DQ under the Impartial Culture assumption.

We define the **Dodgson Relaxed (DR) score** $\mathrm{Sc}_{\mathbf{R}}$ as with the Dodgson score, but allow votes to be split into rational fractions. However we require that a be swapped over b at least $F(b, a)$ times. Thus $\mathrm{Sc}_{\mathbf{R}}(d) \geq \mathrm{Sc}_{\mathbf{Q}}(d)$ and the Dodgson Relaxed rule will converge at least as quickly as DQ. The DR rules thus sacrifices independence to cloning of the electorate to be closer to the Dodgson rule than DC.

The **Dodgson Relaxed and Rounded (D&) score** $\mathrm{Sc}_{\&}$ is the DR score rounded up, i.e. $\mathrm{Sc}_{\&}(d) = \lceil \mathrm{Sc}_{\mathbf{R}}(d) \rceil$.

3 Dodgson Linear Programmes

The Dodgson Clone scores can be computed by relaxing the integer constraints on the Integer Linear Programme (ILP) for Dodgson's rule (Rothe et al., 2003). In this section we will

show that the difference between the solutions of the ILP and LP are $\mathcal{O}(m!)$. Bartholdi et al. (1989) note that there are only $m!$ orderings of the alternatives and thus no more than $m!$ vote types. However, since we never swap d down the profile (M$^{\text{c}}$Cabe-Dansted, 2006), the ordering of the candidates below d are irrelevant. We will formalise this notion as d-equivalence:

Definition 3.1. *Where* $\mathbf{v}$ *is a linear order on* m *alternatives, let* $\mathbf{v}_i$ *represent the* i^{th} *highest ranked alternative and* $\mathbf{v}_{\leq i}$ *represent the sequence of* i^{th} *highest ranked alternatives. Where* d *is an alternative, we say* $\mathbf{v}$ *and* $\mathbf{w}$ *are* d*-equivalent (*$\mathbf{v} \equiv_d \mathbf{w}$*) iff there exists* i *such that* $\mathbf{v}_i = d$ *and* $\mathbf{v}_{\leq i} = \mathbf{w}_{\leq i}$.

Lemma 3.2. *Let* S_d *be the set of* d*-equivalence classes. Then* $|S_d|$ *is less than* $(m-1)!e$ *where* $e = 2.71\ldots$ *is the exponential constant.*

Proof. We see that there is one equivalence class where d is ranked in the top position, $m-1$ equivalence classes where d is ranked in the second highest position, and in general $\prod_{k=m-i+1}^{m-1} k$ when d is ranked i^{th} from the top. We note that:

$$\prod_{k=m-i+1}^{m-1} k = \frac{(m-1)!}{(m-i)!}$$

We see that

$$\begin{aligned} |S_d| =& \frac{(m-1)!}{(m-m)!} + \frac{(m-1)!}{(m-(m-1))!} + \cdots + \frac{(m-1)!}{(m-1)!} \\ <& (m-1)!\left(\frac{1}{0!} + \frac{1}{1!} + \cdots\right) = (m-1)!e \end{aligned}$$

□

Corollary 3.3. *If we categorise votes into type based on* d*-equivalence classes (instead of linear orders), the ILP below has less than* $m\,(m-1)!e = m!e$ *variables.*

Note that there are less than $(m-1)!e$ choices for i, no more than m choices for j and thus less than $m\,(m-1)!e = m!e$ variables (each of the form y_{ij}).

Lemma 3.4. *We can transform the ILP of Bartholdi et al. (1989) into the following form (M*$^{\text{c}}$*Cabe-Dansted, 2008, 2006):*
$\min \sum_i \sum_{j>0} y_{ij}$ *subject to*
$y_{i0} = N_i$ *(for each type of vote* i*)*
$\sum_{ij}(e_{ijk} - e_{i(j-1)k})y_{ij} \geq D_k$ *(for each alternative* k*)*
$y_{ij} \leq y_{i(j-1)}$ *(for each* i *and* $j > 0$*)*
$y_{ij} \geq 0$*, and each* y_{ij} *must be integer.*

Proof. For each i and j variable y_{ij} represents the number of times that the candidate d is swapped up *at least* j positions. In the LP D_k may be defined as $\text{adv}(k,d)/2$ to compute the DC score or $\lceil \text{adv}(k,d)/2 \rceil$ to compute the DR score (under the ILP these are equivalent). □

Theorem 3.5. *The DR (*$Sc_{\mathbf{R}}$*), DC (*$Sc_{\mathbf{C}}$*) and D&* ($Sc_{\&}$*) scores are bounded as follows:*

$$Sc_{\mathbf{D}}(d) - (m-1)!(m-1)e < Sc_{\mathbf{C}}(d) \leq Sc_{\mathbf{R}}(d) \leq Sc_{\&} \leq Sc_{\mathbf{D}}(d)$$

Proof. Every solution to the Integer Linear Program for the Dodgson score is a solution to the Linear Program for the DR score. Every solution to the LP for the DR score is a solution to the LP for the DC score. Thus the DC score cannot be greater than the DR

score, which cannot be greater than the Dodgson score ($\mathrm{Sc}_{\mathbf{C}}(d) \leq \mathrm{Sc}_{\mathbf{R}}(d) \leq \mathrm{Sc}_{\mathbf{D}}(d)$). Since $\mathrm{Sc}_{\mathbf{D}}$ is integer, it follows that $\mathrm{Sc}_{\mathbf{R}}(d) \leq \lceil \mathrm{Sc}_{\mathbf{R}}(d) \rceil = \mathrm{Sc}_{\&}(d) \leq \mathrm{Sc}_{\mathbf{D}}(d)$. Also note that given a solution y to either LP, we can produce a solution y' to the ILP simply by rounding up each variable ($y'_{ij} = \lceil y_{ij} \rceil$), hence

$$\mathrm{Sc}_{\mathbf{D}}(d) - \mathrm{Sc}_{\mathbf{C}}(d) \leq \sum_i \sum_{j>0} \lceil y_{ij} \rceil - y_{ij} < \sum_i \sum_{j>0} 1 \leq (m-1)!e$$

Since i can take less than $(m-1)!e$ values, and j can vary from 1 to $(m-1)$, it follows that

$$\mathrm{Sc}_{\mathbf{D}}(d) - (m-1)!(m-1)e < \mathrm{Sc}_{\mathbf{C}}(d) \leq \mathrm{Sc}_{\mathbf{R}}(d) \leq \mathrm{Sc}_{\&} \leq \mathrm{Sc}_{\mathbf{D}}(d) \qquad \square$$

These results can also be used to find tighter bounds on the complexity of solving the ILP and LPs (M$^{\mathrm{c}}$Cabe-Dansted, 2006, 2008).

4 Counting Proof of Convergence under IAC

For a voting situation U and linear order $\mathbf{v}$, we represent the number of linear orders of type $\mathbf{v}$ in U by $\#_U(\mathbf{v})$.

Where $X \in \{D, T\}$, let $\Delta_{\mathbf{X}}(a, z)$ be equivalent to $\mathrm{Sc}_{\mathbf{X}}(a) - \mathrm{Sc}_{\mathbf{X}}(z)$. Given an arbitrary pair of alternatives (a, z) we pick an arbitrary linear order $ab \ldots z$ with a ranked first and z ranked last and call it $\mathbf{v}$. We also define the reverse linear order $\tilde{\mathbf{v}} = z \ldots ba$.

Lemma 4.1. *Replacing a vote of type $\tilde{\mathbf{v}}$ with a vote of type $\mathbf{v}$ will increase $\Delta_{\mathbf{T}}(a, z)$ by at least one.*

Proof. We see that replacing a vote of type $\tilde{\mathbf{v}}$ with a vote of type $\mathbf{v}$ will increase $\mathrm{adv}(a, z)$ by one, or decrease $\mathrm{adv}(z, a)$ by one. $\square$

Lemma 4.2. *Replacing a vote of type $\tilde{\mathbf{v}}$ with a vote of type $\mathbf{v}$ will increase $\Delta_{\mathbf{D}}(a, z)$ by at least one.*

Proof. Say a is not a Condorcet-tie winner, but is a Condorcet-tie winner after some minimal set S of swaps is applied to the profile P. Let P' be the profile P after one vote of type $\tilde{\mathbf{v}}$ has been replaced with a vote of type $\mathbf{v}$. If any swaps were applied to the vote those swaps are no longer required, and $\mathrm{Sc}_{\mathbf{D}}[P'](a) < \mathrm{Sc}_{\mathbf{D}}[P](a)$. Otherwise we can apply the set of swaps S to P' resulting in $\mathrm{adv}(a, k)$ being at least 2 for all other alternatives k. Hence we can remove one of the swaps, and still result in a being a Condorcet-tie winner after the swaps have been applied to P'.

Say a is a Condorcet-tie winner in P. Then z is not a Condorcet-tie winner in P'. As in the previous paragraph we can conclude that $\mathrm{Sc}_{\mathbf{D}}[P](z) < \mathrm{Sc}_{\mathbf{D}}[P'](z)$. $\square$

Lemma 4.3. *For a fixed integer k, and a fixed ordered pair of alternatives (a, z) the proportion of voting situations, with n agents and m alternatives, for which $\Delta_{\mathbf{X}}(a, z) = k$ is no more than:*

$$\frac{(m! - 2)}{(n + m! - 2)}$$

Proof. Say $\mathbf{v} = ab \ldots z$ is some fixed linear order and $\tilde{\mathbf{v}} = z \ldots ba$ is the reverse order. We define an equivalence relation $\sim$ on the set of voting situations, as follows: say U, V are two voting situations, then

$$U \sim V \iff \forall_{\mathbf{w} \neq \mathbf{v},\, \mathbf{w} \neq \tilde{\mathbf{v}}} \#_V(\mathbf{w}) = \#_U(\mathbf{w}).$$

From Lemma 4.1 and 4.2 we see that in each equivalence class, there can be at most one voting situation for which $\Delta_{\mathbf{X}}(a,z) = k$. Also note that whereas there are

$$|\mathcal{S}^n(A)| = \binom{n+m!-1}{n}$$

distinct voting situations there are at most

$$\binom{n+(m!-1)-1}{n}$$

equivalence classes under $\sim$. Hence the proportion of voting situations for which $\Delta_{\mathbf{X}}(a,z) = k$ is no more than:

$$\frac{(n+m!-1)!}{n!(m!-1)!}\frac{n!(m!-2)!}{(n+m!-2)!} = \frac{(n+m!-1)!}{(n+m!-2)!}\frac{(m!-2)!}{(m!-1)!} = \frac{(m!-2)}{(n+m!-2)} \qquad \square$$

Lemma 4.4. *If $\Delta_{\mathbf{T}}(a,z) = k$ and a is a Tideman winner and z is a DQ winner, then $0 \leq k < m$.*

Proof. As a is a Tideman winner and z is a DQ winner, then

$$\mathrm{Sc}_{\mathbf{T}}(a) \leq \mathrm{Sc}_{\mathbf{T}}(z), \quad \mathrm{Sc}_{\mathbf{Q}}(z) \leq \mathrm{Sc}_{\mathbf{Q}}(a)$$

Recall that

$$\mathrm{Sc}_{\mathbf{Q}}(x) = \sum_{y \neq x} \left\lceil \frac{\mathrm{adv}(y,x)}{2} \right\rceil, \quad \mathrm{Sc}_{\mathbf{T}}(x) = \sum_{y \neq x} \mathrm{adv}(y,x).$$

We see that $\mathrm{adv}(y,x) \leq 2\lceil \mathrm{adv}(y,x)/2 \rceil \leq \mathrm{adv}(y,x) + 1$ and so $\mathrm{Sc}_{\mathbf{T}}(x) \leq 2\mathrm{Sc}_{\mathbf{Q}}(x) < \mathrm{Sc}_{\mathbf{T}}(x) + m$ for all alternatives x. Thus

$$\mathrm{Sc}_{\mathbf{T}}(z) \leq 2\mathrm{Sc}_{\mathbf{Q}}(z) \leq 2\mathrm{Sc}_{\mathbf{Q}}(a) < \mathrm{Sc}_{\mathbf{T}}(a) + m.$$

And so $\mathrm{Sc}_{\mathbf{T}}(a) \leq \mathrm{Sc}_{\mathbf{T}}(z) < \mathrm{Sc}_{\mathbf{T}}(a) + m$. Let $k = \mathrm{Sc}_{\mathbf{T}}(a) - \mathrm{Sc}_{\mathbf{T}}(z)$. Then $0 \leq k < m$, and so there are no more than m ways of choosing k if we wish the DQ and Tideman winners to differ. $\square$

Recall that Theorem 3.5 states:

$$\mathrm{Sc}_{\mathbf{D}}(d) - (m-1)!(m-1)e < \mathrm{Sc}_{\mathbf{C}}(d) \leq \mathrm{Sc}_{\mathbf{R}}(d) \leq \mathrm{Sc}_{\&} \leq \mathrm{Sc}_{\mathbf{D}}(d)$$

where $e = 2.71\ldots$ is the exponential constant. Given that there are only m ways of choosing k such that the Tideman and DQ winners differ, and less than $(m-1)!(m-1)e$ ways of choosing k such that the Dodgson, DC, DR and/or D& winners differ, we get the following theorem.

Theorem 4.5. *The proportion of voting situations, with n agents and m alternatives, for which a is a Tideman winner and z is a DQ winner is no more than:*

$$\frac{(m!-2)}{(n+m!-2)}m.$$

The proportion for which a is Dodgson winner and z is a DC, DR and/or D& winner is less than

$$\frac{(m!-2)}{(n+m!-2)}(m-1)!(m-1)e.$$

As there are only $m(m-1)$ ways of choosing a and z from the set of alternatives, we get the following corollary

Corollary 4.6. *The probability that the DQ and Tideman rule pick the same winners converges to* 1 *as* $n \to \infty$*, under the Impartial Anonymous Culture assumption. Likewise the probability that the DR, DC, D& and Dodgson rules pick the same winner converges to 1 as* $n \to \infty$*.*

In other words the DR, DC and D& winners (and scores) provide "deterministic heuristic polynomial time" (Procaccia and Rosenschein, 2007) algorithms for the Dodgson winner (and score) with the IAC distribution.We can use the same technique to show that the Greedy Algorithm proposed by Homan and Hemaspaandra (2005) converges to the DQ and Tideman rules under IAC, as the `GreedyScore` differs from the DQ score by less than m.By setting k to 0 we may likewise prove that the probability of a non-unique Tideman or Dodgson winner converges to 0 under IAC.

We extend the concept of a "frequently self-knowingly correct algorithm" (Homan and Hemaspaandra, 2005) such that we can specify a distribution over which the algorithm is frequently self-knowingly correct.

Definition 4.7. *A self-knowingly correct (Homan and Hemaspaandra, 2005) algorithm* A *is a "frequently self-knowingly correct algorithm over a distribution* μ*" for* $g\colon \Sigma^* \to T$ *iff*

$$\lim_{n\to\infty} \sum_{x\in\Sigma^n,\, A(x)\in T\times\{maybe\}} P_\mu(x) = 0$$

where $P_\mu(x)$ *is the probability that* $X = x$ *when* X *is chosen from* Σ^n *under the* μ *distribution and for all* x *we have* $(A(x))_1 = g(x)$ *or* $(A(x))_2 =$ *"maybe".*

Using the set of linear orders $\mathcal{L}(\mathcal{A})$ as Σ we may construct a frequently self-knowingly correct algorithm from DC (or DR, or D&) as the algorithm can output "definitely" whenever the DC winner has DC score that is at least $(m-1)!(m-1)e$ less than any other alternative.

5 Non-convergence of Tideman Based rules

Definition 5.1. *A "voting ratio" is a function* $f\colon \mathcal{L}(\mathcal{A}) \to [0,1]$ *such that*

$$\sum_{\mathbf{v}\in\mathcal{L}(\mathcal{A})} f(\mathbf{v}) = 1.$$

We say that a profile $\mathcal{P}$ *reduces to a voting ratio* f *if*

$$\forall_{\mathbf{v}\in\mathcal{L}(\mathcal{A})} \#\{i : \mathcal{P}_i = \mathbf{v}\} = nf(\mathbf{v})$$

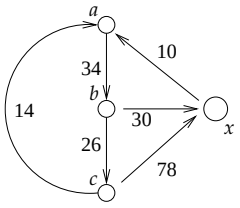
Note 5.2. *A voting ratio is similar to a voting situation, but unlike a voting situation does not contain any information about the total number of agents.*

Definition 5.3. *We say that a voting ratio* f *is "bad" if for every profile* $\mathcal{P}$ *that reduces to* f *and has an even number of agents, the DQ winner of* $\mathcal{P}$ *differs from the Dodgson winner.*

Example 5.4. *The following voting ratio is bad.*

$$h(\mathbf{v}) = \begin{cases} {}^{16}/_{39} \text{ if} & \mathbf{v} = abcx \\ {}^{12}/_{39} \text{ if} & \mathbf{v} = cxab \\ {}^{10}/_{39} \text{ if} & \mathbf{v} = bcxa \\ {}^{1}/_{39} \text{ if} & \mathbf{v} = cbax \\ 0 & \text{otherwise} \end{cases}$$

We see that for a profile that reduces to the above voting ratio, we have the following advantages and scores per 78 agents:



	a	b	c	x
DQ score	12	17	13	54
Dodgson score	14	17	13	54

The Dodgson score of a is higher than the DQ score of a as to swap a over c we must first swap a over b or a over x. We have to swap a over c at least 7 times. Hence we must use a total of at least 14 swaps to make a a Condorcet-tie winner. We see that a is the DQ winner, but c is the Dodgson winner.

We will now show that there exists a neighbourhood around the voting ratios g and h that is bad.

Lemma 5.5. *Altering a single vote will change the Dodgson scores and DQ scores by at most $m-1$.*

Proof. Recall that

$$\mathrm{Sc}_{\mathbf{Q}}(a) = \sum_{b \neq a} F(b,a), \text{ where } F(b,a) = \left\lceil \frac{\mathrm{adv}(b,a)}{2} \right\rceil .$$

We see that for each other alternative b changing a single vote can change $F(b,a)$ by at most one, and there are $m-1$ such alternatives.

Say that P and R are two profiles that differ only in a single vote. Let P' and R' be P and R respectively after some arbitrary d has been swapped to the top of the vote that differs, which requires no more than $m-1$ swaps. We see that

$$\begin{aligned} &\mathrm{Sc}_{\mathbf{D}}[P](d) \leq \mathrm{Sc}_{\mathbf{D}}[P'](d) + m - 1 \leq \mathrm{Sc}_{\mathbf{D}}[R](d) + m - 1 \\ &\mathrm{Sc}_{\mathbf{D}}[R](d) \leq \mathrm{Sc}_{\mathbf{D}}[R'](d) + m - 1 \leq \mathrm{Sc}_{\mathbf{D}}[P](d) + m - 1 \ . \end{aligned}$$

□

Corollary 5.6. *For any positive integer k, alternative d, profile P and rule $X \in \{D, Q\}$ (i.e. Dodgson or DQ), if $Sc_{\mathbf{X}}(d) < Sc_{\mathbf{X}}(a) - 2k(m-1)$ for all other alternatives a, then d will remain the unique X winner in any profile that results from changing k or less votes.*

Lemma 5.7. *There is a neighbourhood of bad voting ratios around the voting ratio h from Example 5.4.*

Proof. We see that in a profile with n agents that reduces to the voting ratio h the Dodgson and DQ winners have scores that are at least $\frac{n}{78}$ lower than the other alternatives. Thus if we alter less than $\frac{n}{78}\frac{1}{2(1)(4-1)} = \frac{n}{468}$ votes, the DQ winner will remain different from the Dodgson winner. □

Thus we may now state the following theorem:

Theorem 5.8. *If we generate profiles randomly according to the IAC distribution, with $m \geq 4$, the DQ, Greedy and Tideman winners do not converge to the Dodgson winner as the number of agents tends to infinity.*

It is easy to generalise this result to non-trivial Pólya-Eggenberger distributions. See (M^{c}Cabe-Dansted, 2006) for details.

Under the IAC, the Tideman rule (M^{c}Cabe-Dansted et al., 2007) and the greedy algorithm proposed by Homan and Hemaspaandra (2005) converges to the DQ rule, which does not converge to Dodgson's rule. It follows that the Tideman rule and greedy algorithm do not converge to Dodgson's rule under the IAC.

As the difference between the DQ scores and the `GreedyScores` (Homan and Hemaspaandra, 2005) is less than m, the DQ winner and `GreedyWinner` will converge under the IAC.

6 Conclusion

We have previously (M^{c}Cabe-Dansted et al., 2007) proved that Tideman's rule does converge under IC, and presented a refined rule (DQ) which converged exponentially quickly. Another approximation with exponentially fast convergence was independently proposed by Homan and Hemaspaandra (2005).

Unfortunately, real voters are not independent, and so the Impartial Culture assumption is not realistic. This paper investigates the asymptotic behaviour of approximations to the Dodgson rule under other assumptions of voting behaviour, particularly the Impartial Anonymous Culture (IAC) assumption, that each multiset of votes is equally likely.

We found that the approximations converge to each other under IAC, but they do not converge to Dodgson. Hence the DQ rule is not asymptotically closer to the Dodgson rule than the Tideman rule is, although DQ converges faster under the IC.

It is not realistic to assume that voters' behaviour will precisely follow any mathematical model, including IAC. Fortunately the proof that our new approximations converge to Dodgson does not depend on the details of the IAC. Indeed, the DC, DR, D& and Dodgson winners will all be the same if the scores of the two leading alternatives differ by less than $(m-1)!(m-1)e$. This assumption is realistic for a sufficiently large number of voters. Even in the neck and neck 2000 US presidential elections, the popular vote for the leading two alternatives differed by half a percent — over half a million votes. If we used the Dodgson rule to choose between the top five alternatives, the difference between the Dodgson scores of the leading two alternatives would have to be less than 261 for the winners to differ. Even if the entire electorate conspired to cause the winners to differ, minor inaccuracies such as hanging chads could frustrate their attempt. Even in cases where the Dodgson Relaxed does differ from the Dodgson, the difference seems to be primarily that the Dodgson Relaxed rule picks a smaller set of tied winners as it is able to split ties in favour of alternatives that have fractionally better DR scores. This is in some sense more democratic than tie breaking procedures such as breaking ties in favour of the preferences of the first voter.

Thus determining that an algorithm is "frequently self-knowingly correct", as defined by Homan and Hemaspaandra (2005), is insufficient to conclude that the algorithm will converge in practice. The DQ and `GreedyWinner` provide frequently self-knowingly correct algorithms, but do not converge under other assumptions of voter behaviour such as IAC. We have extended the definition to allow a distribution to be specified. So, in other words, DQ and `GreedyWinner` provide algorithms that are "frequently self-knowingly correct over IC", but unlike DC, DR, and D&, do not provide algorithms that are "frequently self-knowingly correct algorithms over IAC".

We have previously shown that the Dodgson scores and winners of a voting situation can be computed from a voting situation with $\mathcal{O}(f_1(m)\ln n)$ arithmetic operations of $\mathcal{O}(f_2(m)\ln n)$ bits of precision (M^{c}Cabe-Dansted, 2006) for some pair of functions f_1 and f_2. However we did not find a good upper bound on f_1 or f_2, even $f_1(4)$ may be unreason-

ably large. This suggests that it may be important to use a variant of Dodgson's rule that is truely polynomial, such as DC, DR or D&. See (M^{c}Cabe-Dansted, 2006) for a discussion of the complexity of these rules.

We have found that the scores of the most of the approximations we have studied form a hierarchy of increasingly tight lower bounds on the Dodgson score:

$$\frac{\mathrm{Sc}_{\mathbf{S}}(x)}{2} \leq \frac{\mathrm{Sc}_{\mathbf{T}}(x)}{2} \leq \mathrm{Sc}_{\mathbf{Q}} \leq \mathrm{Sc}_{\mathbf{R}} \leq \mathrm{Sc}_{\&} \leq \mathrm{Sc}_{\mathbf{D}} \leq \mathrm{Sc}_{\mathbf{R}} + (m-1)!(m-1)e$$

The Dodgson Clone rule does not fit in that hierarchy, although it is the case that

$$\frac{\mathrm{Sc}_{\mathbf{T}}(x)}{2} \leq \mathrm{Sc}_{\mathbf{C}} \leq \mathrm{Sc}_{\mathbf{R}} \leq \mathrm{Sc}_{\mathbf{D}} \leq \mathrm{Sc}_{\mathbf{C}} + (m-1)!(m-1)e \ .$$

Despite the great accuracy of the D& approximation, there are good reasons to pick other approximations. The difference between the DR rule and the D& rule is that the DR rule can split ties based on fractional scores, so the DR rule may be considered superior to D& and Dodgson's rule. The DC rule is resistant to cloning of the electorate. The DQ rule is very simple to compute, and very easy to write in any programming language. As the DQ rule is known to converge exponentially fast under IC, this makes the DQ rule very appropriate for cases where the data is known to distributed according to IC. This is the case for studies that have randomly generated data according to the IC (see e.g. M^{c}Cabe-Dansted and Slinko, 2006; Shah, 2003; Nurmi, 1983). The Tideman rule is no more easy to compute than the DQ rule. However, the mathematical definition of the Tideman rule is simpler than the DQ rule. This makes the Tideman rule useful for theoretical studies of the Dodgson rule where the speed of convergence is not important. Also, like the DC rule, the Tideman rule is resistant to cloning the electorate.

We conclude that the DC and DR rules are superior, for social choice, to the traditional definition of the Dodgson rule. DC provies resistance to cloning the electorate, and both are better at splitting ties than the traditional definition of Dodgson rule. We know of no advantage of the traditional definition over these rules. Even if the traditional Dodgson winner is preferred, it may be hard to justify the computational complexity of the traditional Dodgson rule, especially since it is almost certain that these rules would pick the same result. If, despite all this, the traditional Dodgson rule is still chosen, these new rules provide frequently self-knowingly correct algorithms for the Dodgson winner for any reasonable assumption of voter behaviour, including IAC.

References

Bartholdi, III., Tovey, C. A., and Trick, M. A. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare: Springer-Verlag*, 6:157–165, 1989.

Betzleri, N., Guo, J., and Niedermeier, R. Parameterized computational complexity of dodgson and young elections. In *Proceedings of the 11th Scandinavian Workshop on Algorithm Theory (SWAT'08), Gothenburg, Sweden.*. LNCS, 2008. `http://theinf1.informatik.uni-jena.de/publications/swat_stamped.pdf`.

Black, D. *Theory of committees and elections*. Cambridge University Press, Cambridge, 1958.

Dodgson, C. L. *A method for taking votes on more than two issues*. Clarendon Press, Oxford, 1876. Reprinted in (Black, 1958) with discussion.

Fishburn, P. C. Condorcet social choice functions. *SIAM Journal on Applied Mathematics*, 33:3:469–489, 1977.

Hemaspaandra, E., Hemaspaandra, L., and Rothe, J. Exact analysis of Dodgson elections: Lewis Carroll's 1876 voting system is complete for parallel access to NP. *Journal of the ACM*, 44(6):806–825, 1997.

Homan, C. M. and Hemaspaandra, L. A. Guarantees for the success frequency of an algorithm for finding Dodgson-election winners. Technical Report Technical Report TR-881, Department of Computer Science, University of Rochester, Rochester, NY, 2005. `https://urresearch.rochester.edu/retrieve/4794/tr881.pdf`.

M^{c}Cabe-Dansted, J. C. *Feasibility and Approximability of Dodgson's rule*. Master's thesis, Auckland University, 2006. `http://hdl.handle.net/2292/2614`.

M^{c}Cabe-Dansted, J. C. Dodgson's rule: Approximations and absurdity. Technical report, 2008. `http://www.csse.uwa.edu.au/~john/papers/DodgsonCOMSOC08full.pdf`.

M^{c}Cabe-Dansted, J. C., Pritchard, G., and Slinko, A. Approximability of dodgson's rule. *Social Choice and Welfare*, 2007. (online first) `http://dx.doi.org/10.1007/s00355-007-0282-8`.

M^{c}Cabe-Dansted, J. C. and Slinko, A. Exploratory analysis of similarities between social choice rules. *Group Decision and Negotiation*, 15:1–31, 2006. `http://dx.doi.org/10.1007/s00355-005-0052-4`.

Nurmi, H. Voting procedures: A summary analysis. *British Journal of Political Science*, 13(2):181–208, 1983.

Procaccia, A. D., Feldman, M., and Rosenschein, J. S. Approximability and inapproximability of dodgson and young elections. Discussion Paper Series dp466, Center for Rationality and Interactive Decision Theory, Hebrew University, Jerusalem, 2007.

Procaccia, A. D. and Rosenschein, J. S. Junta distributions and the average-case complexity of manipulating elections. *J. Artif. Intell. Res. (JAIR)*, 28:157–181, 2007.

Rothe, J., Spakowski, H., and Vogel, J. Exact complexity of the winner problem for young elections. *Theory Comput. Syst.*, 36(4):375–386, 2003.

Shah, R. *Statistical Mappings of Social Choice Rules*. Master's thesis, Stanford University, 2003.

Tideman, T. N. Independence of clones as a criterion for voting rules. *Social Choice and Welfare*, 4:185–206, 1987.

Young, H. P. Extending condorcet's rule. *Journal of Economic Theory*, 16:335–53, 1977. `http://dx.doi.org/10.1016/0022-0531(77)90012-6`.

John Christopher M^{c}Cabe-Dansted
M002 The University of Western Australia
35 Stirling Highway, Crawley 6009
Western Australia
Email: `john@csse.uwa.edu.au`

The Cost and Windfall of Manipulability

Abraham Othman and Tuomas Sandholm

Abstract

A mechanism is manipulable if it is in agents' best interest to misrepresent their private information (lie) to the center. We provide the first formal treatment of the *windfall of manipulability*, the seemingly paradoxical quality by which the failure of any agent to play their best manipulation yields a strictly better result than an optimal truthful mechanism. We dub such mechanisms *manipulation optimal*. We prove that any manipulation-optimal mechanism can have at most one manipulable type per agent. We show the existence of manipulation-optimal multiagent mechanisms with the goal of social welfare maximization, but not in dominant strategies when agents are anonymous and the mechanism is symmetric, the most common setting. For this setting, we show the existence of manipulation-optimal mechanisms when the goal is affine welfare maximization.

1 Introduction

Mechanism design is the science of generating rules of interaction—such as auctions and voting protocols—so that desirable outcomes result despite the participation of self-interested agents. A mechanism receives a set of preferences (i.e. type *revelations*) from the agents, and based on that information imposes an *outcome* (such as a choice of president, an allocation of items, and potentially also payments).

A central concept in mechanism design is *truthfulness*, which means that an agent's best strategy is to reveal its type (private information) truthfully to the mechanism. The *revelation principle*, a foundational result in mechanism design, proves that any social choice function that can be implemented in some equilibrium form, can also be implemented using a mechanism where all the agents are motivated to tell the truth. The proof is based on the idea of supplementing the manipulable mechanism with a strategy formulator for each agent that acts strategically on the agent's behalf (see, e.g., Mas-Colell et al. (1995)). Since truthfulness is certainly worth something—with real people, fairness and simplicity, with virtual agents, the elimination of the need to strategically compute—the revelation principle produces something for nothing, a free lunch. As a result, contemporary research into mechanism design has focused almost exclusively on truthful mechanisms.

But manipulable mechanisms protected by the "shield" of computational hardness are intuitively appealing. Computational complexity could be used to sever the symmetry between manipulable and truthful mechanisms, opening up exciting new possibilities in the outcome space. One notable caveat in this agenda is that an agent's inability to find its optimal manipulation does not imply that the agent will act truthfully. Unable to solve the hard problem of finding their optimal manipulation, an agent may submit their true private type but they could also submit their best guess for what their optimal manipulation might be or, by similar logic, give an arbitrary revelation. A challenge in manipulable mechanisms is that it is difficult to predict in which specific ways agents, particularly human agents, will behave if they do not play according to game-theoretic rationality.

In manipulable mechanisms, there are several reasons why agents may fail to play their optimal manipulations. Humans may play suboptimally due to cognitive limitations and other forms of incompetence. The field of behavioral game theory studies the gap between theoretical optimality and human actions (see Camerer (2003) for a survey). Virtual agents may be unable to find their optimal manipulations due to computational lim-

its: finding an optimal strategy is NP-hard in many settings (e.g., Bartholdi et al. (1989); Conitzer and Sandholm (2003, 2004); Procaccia and Rosenschein (2007)), and can be #P-hard (Conitzer and Sandholm, 2003), PSPACE-hard (Conitzer and Sandholm, 2003), or even uncomputable (Nachbar and Zame, 1996). The idea of using complexity as a shield against manipulations has been most prominent in the field of voting theory (e.g., Bartholdi et al. (1989); Conitzer and Sandholm (2003); Procaccia and Rosenschein (2007)).

In this paper, we explore mechanism design beyond the realm of truthful mechanisms using a concept we call *manipulation optimality*, where a mechanism benefits—and does better than any truthful mechanism—if agents fail to play their optimal manipulations *in any way*. This enables the mechanism designer to do better than the revelation principle would suggest, and obviates the need for predicting agents' irrational behavior. Conitzer and Sandholm (2004) show the existence of such a mechanism in an artificial setting, but leave open the question of how broadly this paradigm applies and whether it applies to any practical settings. These are the questions that we answer in this paper.

We prove an impossibility result that curtails the windfall of manipulability significantly. Specifically, we show impossibility if any agent has more than one manipulable type. Curtailed by this first impossibility result, we proceed to study settings where each agent has at most one manipulable type. For single-agent settings, we show impossibility under the social welfare maximization objective and possibility under affine welfare maximization. In contrast, in the multiagent setting we get possibility under both of those objectives, but only under the affine version if agents are symmetric.

2 The general setting

Each agent i has type θ_i and a utility function $u_i^{\theta_i}(o)$, which depends on the outcome o that the mechanism selects. An agent's type captures all of the agent's private information. For brevity, we sometimes write $u_i(o)$.

The mechanism designer has an objective (which can be thought of as mechanism utility) that he tries to maximize:

$$\mathcal{M}(o) = \sum_{i=1}^{n} \gamma_i u_i(o) + m(o),$$

where $m(\cdot)$ captures the designer's desires unrelated to the agents' utilities. This formalism has three widely-explored objectives as special cases:

- Social welfare: $\gamma_i = 1$ and $m(\cdot) = 0$.
- Affine welfare: $\gamma_i > 0$ and $m(\cdot) \geq 0$.
- Revenue: Let outcome o correspond to agent payments to the mechanism of $\pi_1(o), \ldots, \pi_n(o)$. Fix $\gamma_i = 0$ and $m(o) = \sum_{i=1}^{n} \pi_i(o)$.

An agent's type is *manipulable* if reporting some other type yields higher utility for the agent. That report is the agents' *best response* and is generally conditional on the reports of the other agents. If a certain report is a best response for every possible report of the other agents, it is known as a *dominant strategy*. A mechanism implements a social choice function, a function from agent reports to outcomes.

Two manipulable types are *distinct* if, for some revelation of the other agents, the types have different optimal manipulations which lead to different outcomes.

A mechanism M is *truthful* if each agent's dominant strategy in the mechanism is to reveal her true type. A mechanism M is an *optimal truthful mechanism* if it is not (weakly) Pareto-dominated by any other truthful mechanism.

Now we are ready to introduce the main notion of this paper. We say a mechanism is *manipulation optimal* if, when agents play their optimal strategies, the mechanism utility equals that of the best truthful mechanism, and *any* failure of agents to perform their optimal manipulations yields greater mechanism utility.

We assume that, if an agent's optimal play is to reveal her true type, then she will do so. The mechanism, for instance, can publish which types are truthful, and it can be expected that those agents will behave rationally. With software agents, such behavior can be hard-coded in. On the other hand, agents with manipulable types may not behave optimally; for instance, finding an optimal manipulation can be computationally intractable. It is important to note that we do *not* assume that an agent necessarily tells the truth if it fails to find its optimal manipulation.

We now proceed to formalize this. Let o be the outcome that would arise from all agents playing strategically optimally in some manipulable mechanism $\hat{M}$. We will denote by $\hat{o}$ an outcome in $\hat{M}$ that arises if one or more agents fail to perform their optimal manipulations. Now (using the revelation principle), transform $\hat{M}$ into the truthful mechanism M which, given the true types of agents, yields outcome o.

Definition 1 *We call* $\hat{M}$ manipulation-optimal *if it meets the following characteristics:*

1. *M is an optimal truthful mechanism.*
2. $\forall \hat{o} \neq o, \ \mathcal{M}(\hat{o}) > \mathcal{M}(o)$.

2.1 A general impossibility result

While Conitzer and Sandholm (2004) showed that manipulation-optimal mechanisms do exist, the following result strongly curtails their existence generally.

Proposition 1 *No mechanism satisfies Characteristic 2 of Definition 1 if any agent has more than one distinct manipulable type.*

Proof. Suppose, for contradiction, that f is a social choice function satisfying Characteristic 2. Let agent i with type a have a best-response revelation a', and let agent i with type b have a best-response revelation b'. Fix the plays of the other agents as $\mathbf{x}$, such that $f(a', \mathbf{x}) \neq f(b', \mathbf{x})$. Since a and b are distinct, there must exist such an $\mathbf{x}$.

We first define the following shorthand notation:

$$\begin{aligned} \sum(a') &\equiv \sum_{j \neq i} \gamma_j u_j(f(a', \mathbf{x})) + m(f(a', \mathbf{x})) \\ \sum(b') &\equiv \sum_{j \neq i} \gamma_j u_j(f(b', \mathbf{x})) + m(f(b', \mathbf{x})) \end{aligned}$$

Because f satisfies Characteristic 2, we get the following two inequalities on mechanism utilities—for agent i of type b and agent i of type a, respectively.

$$\begin{aligned} \gamma_i u_i^b(f(b', \mathbf{x})) + \sum(b') &< \gamma_i u_i^b(f(a', \mathbf{x})) + \sum(a') \\ \gamma_i u_i^a(f(a', \mathbf{x})) + \sum(a') &< \gamma_i u_i^a(f(b', \mathbf{x})) + \sum(b') \end{aligned}$$

Because a' and b' are best-response plays for agents of their respective types, $u_i^a(f(a', \mathbf{x})) \geq u_i^a(f(b', \mathbf{x}))$ and $u_i^b(f(b', \mathbf{x})) \geq u_i^b(f(a', \mathbf{x}))$. Thus since $\gamma_i \geq 0$ we have

$$\begin{aligned} \gamma_i u_i^b(f(a', \mathbf{x})) + \sum(a') &\leq \gamma_i u_i^b(f(b', \mathbf{x})) + \sum(a') \\ \gamma_i u_i^a(f(b', \mathbf{x})) + \sum(b') &\leq \gamma_i u_i^a(f(a', \mathbf{x})) + \sum(b') \end{aligned}$$

Combining the first lines of the above two equation blocks yields $\sum(b') < \sum(a')$, while combining the second lines yields $\sum(a') < \sum(b')$, a contradiction. ■

Note that this impossibility result is driven by the strict inequality in Characteristic 2 of Definition 1. Weakening the strict inequality to loose inequality results in a very different conclusion: that the mechanism should have identical utilities for reports of both a' and b'. Taken more generally, replacing the strict inequality with loose inequality yields the result that the outcome from reporting any type which is an optimal manipulation must deliver to the mechanism exactly the same utility.

We argue that strict inequality, and the impossibility it implies, is more appropriate for this setting. When we talk about the "windfall of manipulability", what we are talking about are beneficial results beyond the scope of what truthful mechanisms can reach. That is, we want the mechanism to do better when agents make mistakes, not to be so indifferent to agent inputs that it does not matter agents are making mistakes! Moreover, strict inequality was used by Conitzer and Sandholm (2004), so our results are in keeping with that work.

In the rest of this section we explore mechanisms where each agent can have at most one manipulable type; the above impossibility result precludes the existence of manipulation-optimal mechanisms if the other types are not dominant-strategy truthful.

2.2 Single-agent settings

In this subsection we study settings where there is only one agent reporting their private information. (If there are other agents, their types are assumed to be known.)

Proposition 2 *There exist no single-agent manipulation-optimal mechanisms with the objective of social welfare maximization.*

Proof. In the single-agent context, social welfare maximization means maximizing the utility of the single agent. For contradiction, let f be a manipulation-optimal mechanism. Let the agent of type a have optimal play a' such that $u(f(a')) \geq u(f(x))\ \forall x \in \Theta_i$. But by Characteristic 2, $u(f(a')) < u(f(x))\ \forall x \in \Theta_i \setminus a'$. ■

Proposition 3 *There exist single-agent manipulation-optimal mechanisms with the objective of affine welfare maximization.*

Proof. We can derive this result from the constructive proof of Conitzer and Sandholm (2004). Because the transformation is non-trivial, we restate that result here.

There exists a manager with three possible true types for a team of workers that needs to be assembled:

1. "Team with no friends", which we abbreviate TNF.
2. "Team with friends", which we abbreviate TF.
3. "No team preference", which we abbreviate NT.

The mechanism implements one of two outcomes: picking a team with friends (TF), or picking a team without friends (TNF). The manager gets a base utility 1 if TNF is chosen, and 0 if TF is chosen. If a manager has a team preference, implementing that team preference (either with or without friends) gives the manager an additional utility of 3.

In addition to the manager, the other agent in the game is the HR director, who has utility 2 if a team with friends is chosen. Even though there are two agents in the game, because the HR director does not report a type, this is not a multiagent setting. In fact, the HR director's utilities are equivalent to the payoffs from the outcome-specific mechanism utility map m.

The optimal truthful mechanism maps reports of NT and TNF to TNF and TF to TF. Now consider the manipulable mechanism which maps reports of TNF to TNF and NT and TF to TF. Note that in this mechanism there is only one manipulable type, NT, and that its optimal strategic play is to report TNF. This mechanism is manipulation-optimal: if the manager has type NT and reports NT or TF instead of TNF, the mechanism generates affine welfare of 2, whereas the optimal truthful mechanism generates affine welfare of 1. ■

Conitzer and Sandholm (2004) showed that, for an NT agent, reporting TNF is NP-hard because actually constructing a team of size k without friends requires solving the independent set problem in a graph of people where the edges are friend relationships. Computational complexity is a strong justification for why an agent may not be able to find its optimal manipulation.

2.3 Multi-agent settings

Though we proved above that there do not exist single-agent social welfare maximizing manipulation-optimal mechanisms, they do exist in multi-agent settings.

Proposition 4 *There exist multi-agent manipulation-optimal mechanisms with the objective of social welfare maximization.*

Proof. Consider a game in which two agents, the row agent and the column agent, can have one of two types, a or a'. Our mechanism maps reports to one of four different outcomes:

Report	a'	a
a'	o_1	o_2
a	o_3	o_4

The following two payoff matrices over the four outcomes constitute a manipulation-optimal mechanism. Payoffs for type a are on the left and for type a' on the right:

Report	a'	a
a'	1,1	4,0
a	0,3	3,0

Report	a'	a
a'	3,4	5,0
a	0,6	0,0

Here, playing a' is a strictly dominant strategy for agents of both types. By the revelation principle, we can "box" this mechanism into a truthful mechanism, M_1, that always chooses o_1. However, when an agent of type a plays a rather than a', social welfare is strictly higher than with o_1. (This holds regardless of how others play.) We have now proven Characteristic 2.

What remains to be proven is Characteristic 1; we must demonstrate that M_1 is optimal among truthful mechanisms. We begin by examining the following table, which lists the payoffs for agents over outcomes given the four possible true type combinations:

True types	o_1	o_2	o_3	o_4
a, a	2	4	3	3
a, a'	5	4	6	3
a', a	4	5	3	0
a', a'	7	5	6	0

Now, for contradiction, let M^D be a truthful mechanism that Pareto dominates M_1. Note that M_1 delivers the highest payoff when both agents are of type a'. Thus, $M^D(a', a') = o_1$. But this implies $M^D(a, a')$ and $M^D(a', a)$ must also equal o_1: mapping them to the outcome that gives higher social welfare (in the former case, o_3 and in the latter, o_2) is

not truthful because the agent of type a has incentive to report a' and force o_1. At the same time, mapping to an outcome that is not o_1 delivers less social welfare than M_1. So, $M^D(a', a') = M^D(a', a) = M^D(a, a') = o_1$. But if these three inputs map to o_1, M^D cannot truthfully map revelations of (a, a) to any outcome other than o_1, because some agent will always want to deviate, report type a', and force outcome o_1. Thus, our construction of M^D fails because $M^D = M_1$. ■

The result above uses dominant strategy equilibrium as the solution concept. Therefore, the result implies possibility for weaker equilibrium notions as well.

The agents in our construction are not symmetric. (Symmetry means that all agents have the same payoffs for their reports relative to the reports of the other agents.) We may ask whether manipulation-optimal mechanisms exist for what can be considered the most common setting: where agents are symmetric, the mechanism is anonymous, and the objective is welfare maximization. (By anonymous we mean that the mechanism selects an outcome based only on the distribution of reported types, rather than the agents who reported those types.)

Proposition 5 *There exist no dominant-strategy anonymous multi-agent manipulation-optimal mechanisms with the objective of social welfare maximization for symmetric agents.*

While the impossibility results earlier in this paper were based on a violation of Characteristic 2 of manipulation-optimal mechanisms alone, here the impossibility comes from not being able to satisfy Characteristics 1 and 2 together.

Proof. By Proposition 1, we can focus on mechanisms with a single manipulable type. Call the type a, with dominant strategy a'. Suppose mechanism $\hat{M}$ satisfies Characteristic 2. By the revelation principle it has a corresponding truthful mechanism M. We show that we can construct a truthful mechanism M that Pareto dominates M'.

First, if a set of reports includes a type other than a or a', we set M^D to simply mirror the action taken by M. Strategic implications for agents other than types a and a' are unaffected because for agents of those types revealing their true type was a dominant strategy under $\hat{M}$.

Let o be the outcome implemented by M when all agents reveal a, and let o' be the outcome implemented by M when all agents reveal a'. Denote by $\tilde{a}$ any combination of revelations a and a'; note that $M'(\tilde{a}) = o'$.

By Characteristic 2 we know that we get higher social welfare if agents of type a—whose best manipulation is to report a'—cannot find the manipulation and report a instead. Since agents are symmetric, this implies $u^a(o') < u^a(o)$. This is akin to the Prisoner's Dilemma: the dominant strategy of type a is to report a' , but the outcome is worse for agents if they all report a' rather than a.

Now we construct M^D based on the payoff structure of agents of type a'.

- **Case I:** $u^{a'}(o') < u^{a'}(o)$. In this case we let M^D map each $\tilde{a}$ to o. M^D Pareto dominates M.
- **Case II:** $u^{a'}(o') \geq u^{a'}(o)$. In this case we let M^D select o if all agents report a and o' for any other $\tilde{a}$. M^D Pareto dominates M. Note that M^D is identical to M for all reports except the one where all agents report a. ■

Note that both our possibility results and this impossibility result have used the dominant strategy solution concept. This implies the strongest possibility, but the weakest impossibility. Here, our requirement for dominant strategy manipulability avoids issues with degenerate special cases.

We can get around this impossibility by moving to the affine welfare objective. Note that for an anonymous mechanism, the outcome-specific mechanism utility function $m(\cdot)$

can depend only on the distribution of types, rather than the identities associated with those types.

Proposition 6 *There exist dominant-strategy anonymous multi-agent manipulation-optimal mechanisms with the objective of affine welfare maximization for symmetric agents.*

Proof. We provide a constructive proof with the same framework as Proposition 4. But now let the payoff matrices be as follows (left matrix for type a and right matrix for type a').

Report	a'	a
a'	2,2	1,1
a	1,1	0,0

Report	a'	a
a'	4,4	1,3
a	3,1	0,0

Let $\gamma_i = 1$ for all i, and let the mechanism's additional payoff, $m(\cdot)$, be $\{0, 3, 3, 5\}$ for outcomes o_1 through o_4, respectively. Note that the row and column agents are symmetric (the payoff matrices are symmetric) and that $m(o_2) = m(o_3)$. The dominant strategy equilibrium for this mechanism is for every agent to report type a'. Therefore this mechanism has truthful analogue M_1, the mechanism that always chooses o_1.

We now show that M_1 is an optimal truthful mechanism. First, note that M_1 maximizes the objective when both agents have type a'. It can be shown that (using a construction akin to the last table in the proof of Proposition 4) that due to agent incentives to deviate, any truthful mechanism that would dominate M_1 must map all reports to o_1. Therefore M_1 is an optimal truthful mechanism.

The manipulation-optimality of the mechanism defined by the payoff matrices above comes from noting that whenever agents of type a fail to report a', affine welfare is strictly higher. ■

3 Conclusions and Future Directions

The strategic equivalence of manipulable and truthful mechanisms—captured by the revelation principle—does not mean that any manipulable mechanism is automatically flawed. The failure of agents to perform their best response (or play a particular equilibrium among many), either due to computational constraints or any flavor of incompetence, can actually increase mechanism utility. When the equivalent truthful mechanism to such a manipulable mechanism is optimal among truthful mechanisms, then the manipulable mechanism is truly a better solution. We call such a mechanism *manipulation optimal*.

For a completely general setting, we show that manipulation optimality is limited to mechanisms that have at most one manipulable type per agent. Thus there is a "cost of manipulability" — implementing a manipulable mechanism inherently exposes the designer to achieving an unnecessarily poor result when agents do not perform optimally. This result is, in large part, in line with the revelation principle, although here the considerations are more subtle and the impossibility not universal. This is an inauspicious finding for the concept of using computational complexity as a "trick" to get around the revelation principle: if our mechanism utility function is sufficiently non-trivial (i.e. so that agents' reports with manipulable types can affect it), then the mechanism designer is exposed to the risk of bad outcomes.

It is worth noting how our results apply to the now-copious literature on the complexity of voting schemes. Here they are less disheartening, because non-trivial voting schemes are inherently manipulable by the Gibbard-Satterthwaite impossibility result. So in the voting setting, there is no "truthful analogue" that our manipulable mechanism is performing worse

than. Note also that most voting schemes use cardinal utility (ranked preference lists) as opposed to the ordinal utilities employed here.

It would be interesting to study manipulation optimality under other objectives, such as notions of fairness. As another direction, we plan to explore whether automated mechanism design (Conitzer and Sandholm, 2002) can be used to design manipulation-optimal mechanisms. Given priors over types (and perhaps also over behaviors), it may be possible to ignore incentive compatibility constraints and design manipulable mechanisms that yield higher mechanism utility.

Acknowledgments

This material is based upon work supported by the National Science Foundation under ITR grant IIS-0427858.

References

John Bartholdi, III, Craig Tovey, and Michael Trick. The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6(3):227–241, 1989. ISSN 0176-1714.

Colin Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, 2003.

Vincent Conitzer and Tuomas Sandholm. Complexity of mechanism design. In *Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 103–110, Edmonton, Canada, 2002.

Vincent Conitzer and Tuomas Sandholm. Universal voting protocol tweaks to make manipulation hard. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 781–788, Acapulco, Mexico, 2003.

Vincent Conitzer and Tuomas Sandholm. Computational criticisms of the revelation principle. In *The Conference on Logic and the Foundations of Game and Decision Theory (LOFT)*, Leipzig, Germany, 2004. Earlier versions: AMEC-03, EC-04.

Andreu Mas-Colell, Michael Whinston, and Jerry R. Green. *Microeconomic Theory*. Oxford University Press, New York, NY, 1995.

John H Nachbar and William R Zame. Non–computable strategies and discounted repeated games. *Economic Theory*, 8(1):103–122, June 1996.

Ariel D. Procaccia and Jeffrey S. Rosenschein. Junta Distributions and the Average-Case Complexity of Manipulating Elections. *Journal of Artificial Intelligence Research (JAIR)*, 28:157–181, 2007.

Abraham Othman
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
Email: `aothman@cs.cmu.edu`

Tuomas Sandholm
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
Email: `sandholm@cs.cmu.edu`

Informational requirements of social choice rules

Shin Sato[1]

Abstract

The needed amount of information to make a social choice is the cost of information processing, and it is a practically important feature of social choice rules. We introduce informational aspects into the analysis of social choice rules and prove that (i) if an anonymous, neutral, and monotonic social choice rule operates on minimal informational requirements, then it is a supercorrespondence of either the plurality rule or the antiplurality rule, and (ii) if the social choice rule is furthermore Pareto efficient, then it is a supercorrespondence of the plurality rule.

Keywords: antiplurality rule, minimal informational requirement, plurality rule, social choice rule.

1 Introduction

Each social choice rule utilizes information on the agents' preferences at different levels. For example, it is intuitively clear that dictatorship needs much less information than the Borda rule; under dictatorship, we need to know only the most preferred alternative of a dictator, while under the Borda rule, we need to know the whole preferences of all agents. Without some electronic device (this is the case in most situations where collective choice is to be made)[2], processing a large amount of information is not an easy task. The required amount of information can be considered as the cost of information processing; the larger the amount of information to process, the more time and human resources are needed and the more risk of making errors is involved.

Therefore, the informational requirement is a practically important feature of each social choice rule. That is, when the information processing cost is high, informational requirements should be one of the most important criteria of social choice rules in evaluating them. Therefore, in this paper, we incorporate the informational aspects into social choice. The fundamental problem we are to deal with is the following, "Given a group of social choice rules satisfying some "reasonable" properties, which of them operates on the smallest amount of information?" In other words, we incorporate minimal informational requirements into the axiomatic analysis of social choice rules.

Our main results are (i) if an anonymous, neutral, and monotonic social choice rule operates on minimal informational requirements, then it utilizes only information about either the top ranked alternatives or the bottom ranked alternatives by the agents and it is a supercorrespondence of either the plurality rule or the antiplurality rule, and (ii) if the social choice rule is furthermore efficient, then it utilizes only information about the top ranked alternatives by the agents and it is a supercorrespondence of the plurality rule. Thus, the plurality rule and the antiplurality rule are characterized as the most selective social choice rules among anonymous, neutral, and monotonic social choice rules which operate on minimal informational requirements, and the plurality rule can be characterized as the most selective social choice rule among anonymous, neutral, monotonic, and efficient social choice rules which operate on minimal informational requirements.

This last result is the easiest one to interpret. The plurality rule is widely used in our daily lives, and many people would agree that, compared with other "reasonable" social choice rules, the main

[1]I am grateful to three anonymous referees of this conference for helpful comments, especially, for letting me notice the literature on communication complexity.

[2]In some area of the world, electronic voting systems are adopted in some "big" elections. However, it is very unlikely that all social choices ranging from national elections to the choice of restaurant for a dinner are made with a electronic device, at least in the near future. Major obstacles for electronic voting systems are the cost of introducing the system and the reliability of hardware and software. Actually, in Japan, the result of the election in Kani city in 2003 was cancelled due to a hardware problem, and in Aki ward of Hiroshima city, electronic voting is abandoned in 2006 due to the financial constraint.

advantage of the plurality rule lies in its simplicity and selectivity (i.e., the set of "winners" is small). Our last result theoretically supports this common sense.

Let us mention some related literature. Conitzer and Sandholm (2005) present *communication complexities* of eleven major voting rules.[3] (See Kushilevitz and Nisan (1997) for a survey on the literature on communication complexity. A seminal work is Yao (1979).) In their model, each agent sends a bit of his private information necessary to make a social choice to the others sequentially. That is, the agents "communicate" to compute the value of a voting rule. Communication complexity of a voting rule is defined as the worst-case number of bits in the best protocol to compute the value of the voting rule. Communication complexity can be considered as a kind of informational size of a voting rule. Among many differences, the most significant and essential one between my approach and Kushilevitz and Nisan (1997) is that I introduce a minimal informational requirement as an "axiom" and hence measuring the informational size of some specific social choice rules is not my objective while it is in Kushilevitz and Nisan (1997).

Many social choice rules are proposed and axiomatically characterized in the theory of social choice.[4] Being prevalent in the real world, the plurality rule is axiomatically characterized by Richelson (1978); Roberts (1991); Ching (1996); Yeh (2008), among others. Our contribution to this literature is to characterize the plurality rule (and the antiplurality rule) based on minimal informational requirements and selectivity.

Some researchers consider social choice rules which rely on limited information on preferences. (For example, Moulin (1980); Roberts (1991); Yeh (2008), among others.) However, in their analyses, such restrictions are put as assumptions and do not intend to study the amount of necessary information to make a social choice under each social choice rule.

In sum, analyses in this paper such as investigation of the minimal informational size needed to be a "reasonable" social choice rule and characterizations based on minimal informational requirements seem to be novel in the literature, and would give useful insights in the evaluation of social choice rules.

In Section 2, we give basic notation and definitions. In Section 3, a series of results are presented. Proofs are collected in Section 4.

2 Basic notation and definitions

Let $N = \{1, \ldots, n\}$ be a finite set of agents and let X be a finite set of alternatives with $|X| = m \geq 2$. Let $\mathcal{L}$ denote the set of all linear orders (complete, transitive, and antisymmetric binary relations) on X. An element $R_N = (R_1, \ldots, R_n)$ of $\mathcal{L}^N$ is called a preference profile. A linear order R_i in a preference profile R_N is agent i's preference, and P_i is the strict part of R_i. For each preference $R \in \mathcal{L}$ and for each integer k with $1 \leq k \leq m$, let $r_k(R)$ denote the kth ranked alternative with respect to R. For each $i \in N$, a function φ_i of $\mathcal{L}$ onto a finite set $\mathcal{K}_i$ is called a ***message function*** and a set $\mathcal{K}_i$ is called a ***message space***. A triple $(\varphi_N, \mathcal{K}_N, f)$ is called a ***rule***, where φ_N is a profile of message functions $(\varphi_1, \ldots, \varphi_n)$, $\mathcal{K}_N$ is the Cartesian product of message spaces $\mathcal{K}_i$, and f is a correspondence of $\mathcal{K}_N$ into X. When the agents have a preference profile R_N, then they report a message profile $\varphi_N(R_N) = (\varphi_1(R_1), \ldots, \varphi_n(R_n)) \in \mathcal{K}_N$ and f makes a choice based on the received message $\varphi_N(R_N)$.

For our purpose, the labels or the names of messages are inessential and we restrict the form of message spaces (and message functions) to a specific form without loss of generality. This can be done as follows; let $(\varphi_N, \mathcal{K}_N, f)$ be a rule with a general form.

[3] More precisely, Conitzer and Sandholm (2005) present the asymptotic lower and upper bounds of communication complexities of voting rules. (For example, the plurality rule belongs to $\Theta(n \log_2 m)$.) This is a standard way to measure efficiency of an algorithm in computer science.

[4] See Sen (1986) and Moulin (1988), among others, for surveys of the literature.

- (Message spaces) For each $i \in N$, we can define the partition $\mathcal{M}_i$ of $\mathcal{L}$ induced by φ_i^{-1}. Formally, $\mathcal{M}_i = \{\varphi_i^{-1}(k_i) \mid k_i \in \mathcal{K}_i\}$. Then, we can regard this $\mathcal{M}_i$ as a message space equivalent to $\mathcal{K}_i$ in the sense that there exists a natural bijection between $\mathcal{K}_i$ and $\mathcal{M}_i$; let τ_i be the bijection between $\mathcal{K}_i$ and $\mathcal{M}_i$ defined by $\tau_i(k_i) = \varphi_i^{-1}(k_i)$.

- (Message functions) For each $i \in N$ and for each $R_i \in \mathcal{L}$, let $\varphi_i'(R_i)$ be the element of $\mathcal{M}_i$ such that $R_i \in \varphi_i'(R_i)$. Note that if $\varphi_i(R_i) = k$, then $\tau_i(k) = \varphi_i'(R_i)$. Thus, under φ_i', agent i reports $\varphi_i'(R_i)$, which is a message corresponding to $\varphi(R_i)$. Formally, $\varphi_i' = \tau_i \circ \varphi_i$.

- (Social choice rule) For each $M_N \in \prod_{i \in N} \mathcal{M}_i = \mathcal{M}_N$, let $f'(M_N)$ be $f(k_N)$, where $k_N \in \mathcal{K}_N$ is the message profile corresponding to M_N in the sense that $\tau_N(k_N) = (\tau_1(k_1), \ldots, \tau_n(k_n)) = M_N$. Formally, $f' = f \circ \tau_N^{-1}$.

Now, we have a new rule $(\varphi_N', \mathcal{M}_N, f')$ which is equivalent to $(\varphi_N, \mathcal{K}_N, f)$ in the sense that the only difference is the labels or the names of messages. In $(\varphi_N, \mathcal{K}_N, f)$, agent i reports $\varphi_i(R_i)$. When we just relabel this message $\varphi_i(R_i)$ as $\tau_i(\varphi_i(R_i))$, then we have a rule $(\varphi_N', \mathcal{M}_N, f')$.

Thus, without loss of generality, we can restrict our attention to the rules such that message spaces are partitions of $\mathcal{L}$ and message functions assign each preference the set in the partition to which that preference belongs. In the following, unless otherwise stated, we assume that every rule takes this restricted form.

In the restricted form of rules, a profile of message functions φ_N is uniquely determined by a profile $\mathcal{M}_N$ of message spaces (partitions of $\mathcal{L}$). Thus, in the following, we drop the message functions and write $(\mathcal{M}_N, f)$ for a rule. Given a rule $(\mathcal{M}_N, f)$, when we speak of φ_N, then it should be always understood to be the profile of message functions such that $\varphi_i(R_i) = M_i \in \mathcal{M}_i$ with $R_i \in M_i$. In sum, given a rule $(\mathcal{M}_N, f)$, agents are required to report a profile of sets of linear orders $M_N \in \mathcal{M}_N$ such that the profile of their preferences R_N belongs to M_N, and f makes a choice based on M_N.

It is worth noting that a profile of message spaces $\mathcal{M}_N$ (and hence a profile of message functions φ_N) as well as f is set by the social choice rule designer, and not the variable determined by the agents. We introduce message spaces to clarify what information a social decision requires and to define the informational size of each social choice rule.

Next, we define the informational size of a rule, which is a core concept of this paper.

Definition 2.1 For each rule $(\mathcal{M}_N, f)$, the sum of the numbers of possible messages $\sum_{i \in N} |\mathcal{M}_i|$ is called the *informational size of* $(\mathcal{M}_N, f)$.

Definition 2.2 (The plurality rule) The plurality rule chooses the alternatives ranked as the top by the largest number of agents. In our model, this rule can be written as follows. For each $x \in X$, let $M(x) = \{R \in \mathcal{L} \mid r_1(R) = x\}$. (The set of preferences which rank x at the top.) For each $i \in N$, let $\mathcal{M}_i^p = \{M(x) \mid x \in X\}$. Then, $\mathcal{M}_i^p$ is a partition of $\mathcal{L}$. For each message profile $M_N \in \prod_{i \in N} \mathcal{M}_i^p = \mathcal{M}_N^p$ and for each $x \in X$, let $N_x(M_N) = |\{i \in N \mid M_i = M(x)\}|$. (The number of agents whose message is $M(x)$.) Finally, for each message profile M_N, let $f^p(M_N) = \{x \in X \mid N_x(M_N) \geq N_y(M_N)\ \forall y \in X\}$. Then, $(\mathcal{M}_N^p, f^p)$ is called the *plurality rule*. Its informational size is nm. (Remember that n is the number of the agents and m is the number of alternatives.)

In the plurality rule, f^p makes choice based on information contained in a message profile M_N in $\mathcal{M}_N^p$. From the viewpoint of f^p, it is known that the agent i's preference is in a reported message M_i, but it is not known which is the agent i's preference in M_i. However, the plurality rule can be defined based on this restricted information, because each $M_i \in \mathcal{M}_i^p$ tells what is the alternative ranked as the top.

Thus, in our model, by introducing message spaces between a choice rule and preferences, we can measure the amount of needed information to make a social choice.

Definition 2.3 (The antiplurality rule) The antiplurality rule chooses the alternatives ranked as the bottom by the smallest number of agents. For each $x \in X$, let $M(x) = \{R \in \mathcal{L} \mid r_m(R) = x\}$. (The set of preferences which rank x at the bottom.) For each $i \in N$, let $\mathcal{M}_i^a = \{M(x) \mid x \in X\}$. Then, $\mathcal{M}_i^a$ is a partition of $\mathcal{L}$. For each message profile $M_N \in \prod_{i \in N} \mathcal{M}_i^a = \mathcal{M}_N^a$ and for each $x \in X$, let $N_x(M_N) = |\{i \in N \mid M_i = M(x)\}|$. (The number of agents whose message is $M(x)$.) Finally, for each message profile M_N, let $f^a(M_N) = \{x \in X \mid N_x(M_N) \leq N_y(M_N)\ \forall y \in X\}$. Then, $(\mathcal{M}_N^a, f^a)$ is called the *antiplurality rule*. Its informational size is nm.

At this point, several remarks are in order. First, the reader would notice that to define the informational size and to describe the procedure of making a social choice, it suffices to consider a correspondence which assigns a social outcome to each message profile. For example, in defining the plurality rule, we could define g^p as a correspondence of X^N to X such that for each message profile $x_N = (x_1, \ldots, x_n) \in X^N$, $g^p(x_N)$ is the set of alternatives which are reported by the largest number of agents. If we defined this g^p as the plurality rule, then there would be no agents' "preferences" in the model. The reason to incorporate preferences into our model is that our objective is to find the rules which operate on the minimal information requirements among the rules satisfying some plausible properties such as (weak) monotonicity and efficiency, and these properties refer to agents' preferences. (If our objective were to find the social choice rule which operates on minimal informational requirements without any restriction, then the answer would be constant social choice rules, or "custom", which needs no information to make a social choice.)

Next, although we call φ_i (derived from $\mathcal{M}_i$) a message function and use the word "report", we do not need to interpret them literally. The only role of φ_i is to specify what kind of information is necessary to make a social choice. Thus, we could consider the following model; agents report a preference profile R_N and the central institution which is responsible to make a social decision would take two steps to make a decision. At the first stage, pick up necessary information according to φ_N from R_N, and at the second stage, process information $\varphi_N(R_N) = M_N \in \mathcal{M}_N$ and make a social decision.

Thirdly, we model f as a correspondence and not a function. There are two reasons for this. First, we do not exclude the cases where the society is to choose a set of "satisfactory" alternatives (not necessarily the "best" alternatives). In this case, the social outcome is naturally formulated as sets of alternatives. Second, even when the society is to choose the "best" alternatives, almost all practically important rules such as the plurality rule, the Borda rule, the Copeland rule, and the Simpson rule (See Moulin, 1988), are formulated as correspondences. When we ultimately need to choose a single outcome whereas f can choose multiple alternatives, then it is done by some tie-breaking rule, but this is outside the scope of our analysis.

We define several properties of a rule.

Definition 2.4 A rule $(\mathcal{M}_N, f)$ is said to satisfy

- *anonymity* if for every permutation σ of N and for every $R_N \in \mathcal{L}^N$, $(f \circ \varphi_N)(R_N) = (f \circ \varphi_N)(R_N^\sigma)$, where R_N^σ is defined by for each $i \in N$, $R_i^\sigma = R_{\sigma(i)}$.
- *neutrality* if for every permutation ρ of X and for every $R_N \in \mathcal{L}^N$, $\rho[(f \circ \varphi_N)(R_N)] = (f \circ \varphi_N)[\rho(R_N)]$, where $\rho(R_N) = (\rho(R_1), \ldots, \rho(R_N))$ is defined by for each $i \in N$, $\rho(R_i) = \{(x, y) \in X^2 \mid (\rho^{-1}(x), \rho^{-1}(y)) \in R_i\}$.
- *(weak) monotonicity* if for any R_N and R'_N such that $x \in f(\varphi_N(R_N))$, R_N and R'_N coincide on $(X \setminus \{x\})^2$ and $x = r_k(R_i) = r_{k'}(R'_i)$ with $k' \leq k$ for all $i \in N$, we have $x \in f(\varphi_N(R'_N))$.
- *(Pareto) efficiency* if for any distinct $x, y \in X$ with xP_iy for all $i \in N$, $y \notin f(\varphi_N(R_N))$.

Anonymity requires symmetric treatment of the agents and neutrality requires symmetric treatment of alternatives. Monotonicity requires that when x is chosen at R_N and the position of x (weakly) improves through the change from R_N to R'_N while the relative comparison of any other pair of alternatives is unchanged, then x is still chosen at R'_N. In the literature, it is often called weak monotonicity to distinguish from the so called Maskin monotonicity which does not appear in this paper. Note that monotonicity (in the sense of this paper) is much weaker than the Maskin monotonicity because the relative rankings except x are fixed from the change from R_N to R'_N. Efficiency requires that when an alternative y is dominated by some alternative x, then y cannot belong to the social outcome. Although efficiency is one of the most standard axioms in social choice theory and in economic theory, its relevance depends on the context under consideration. For example, in this paper, as mentioned earlier, we do not exclude cases where the society is to choose a set of "satisfactory" alternatives. In such a case, the fact that y is dominated by x does not imply that y is not satisfactory, and hence y can belong to the social outcome. Based on this observation, we give results with and without efficiency in the next section.

Next, we define formally the minimality of informational requirements.

Definition 2.5 Given a set of rules $\mathcal{F}$, a rule $(\mathcal{M}_N, f)$ is said to *operate on minimal informational requirements in* $\mathcal{F}$ if the informational size of $(\mathcal{M}_N, f)$ is not larger than the informational size of any other rules in $\mathcal{F}$. In this case, the informational size of $(\mathcal{M}_N, f)$ is called the *minimal informational size in* $\mathcal{F}$.

3 Results

In this section, we give a series of results. Let $\mathcal{AN}$ denote the set of nonconstant[5] rules satisfying anonymity and neutrality, let $\mathcal{ANM}$ denote the set of nonconstant rules satisfying anonymity, neutrality, and monotonicity, and let $\mathcal{ANMP}$ denote the set of rules satisfying anonymity, neutrality, monotonicity, and efficiency. Throughout this section, assume $m \geq 2$. (When $m = 1$, then there is no room for "choice".)

Theorem 3.1 *If a rule $(\mathcal{M}_N, f)$ operates on minimal informational requirements in $\mathcal{AN}$, then*

(i) its informational size is nm, and more specifically,

(ii) there exists $h \in \{1, \ldots, m\}$ such that for any $i \in N$, $\mathcal{M}_i = \{M_i(x) \mid x \in X\}$, where $M_i(x) = \{R_i \in \mathcal{L} \mid r_h(R_i) = x\}$.

The second statement of the theorem implies that we can associate each message with one alternative in X and that this relation is a bijection. Moreover, the statement explicitly specifies what information a rule $(\mathcal{M}_N, f)$ depends on; it relies on information what are the hth ranked alternatives in R_N. Consider that agent i with a preference R_i changes his preference to R'_i. Then, agent i sends the same message iff $r_h(R_i) = r_h(R'_i)$.

For example, the plurality rule and the antiplurality rule operate on minimal informational requirements in $\mathcal{AN}$ with $h = 1$ and $h = m$, respectively. Also, the rule $(\mathcal{M}_N, f)$ such that each $\mathcal{M}_i$ is the one defined in the second statement of the theorem with $h = 2$ and f chooses the alternatives second ranked by the largest number of agents also operates on minimal informational requirements in $\mathcal{AN}$.

Before the next theorem, we prepare the following terminology.

Definition 3.1 A rule $(\mathcal{M}_N, f)$ is said to be a *supercorrespondence* of a rule $(\mathcal{M}'_N, f')$ is for every preference profile R_N, $f(\varphi_N(R_N)) \supset f'(\varphi'_N(R_N))$ holds.

[5] A rule $(\mathcal{M}_N, f)$ is said to be *nonconstant* if the correspondence $f \circ \varphi_N$ is nonconstant on $\mathcal{L}^N$.

When a rule $(\mathcal{M}_N, f)$ is a supercorrespondence of a rule $(\mathcal{M}'_N, f')$, then, a rule $(\mathcal{M}_N, f)$ is less selective than $(\mathcal{M}'_N, f')$. Of course, selectivity is not always a plausible axiom. For example, when you want to chooses a set of *satisfactory* (and not necessarily the best) alternatives, then selectivity is not an appealing condition for rules. However, in many cases, we want to choose the socially best alternatives, and in such situations, we usually do not want to rely on a tie-breaking rule (usually, some random device) as much as possible. We want to determine a final social outcome by *preferences* as much as possible. For instance, in elections where we want to choose one winner, it is absurd to use a rule $(\mathcal{M}_N, f)$ such that each voter reports his most preferred candidate and f chooses the candidates who receive at least one vote. (The final outcome is determined by some random device, which is outside our model.)

Theorem 3.2 *If a rule $(\mathcal{M}_N, f)$ operates on minimal informational requirements in $\mathcal{ANM}$, then*

(i) h in Theorem 3.1 is either 1 or m, and

(ii) If $h = 1$, then $(\mathcal{M}_N, f)$ is a supercorrespondence of the plurality rule and if $h = m$, then $(\mathcal{M}_N, f)$ is a supercorrespondence of the antiplurality rule.

This theorem shows that if monotonicity is additionally required, then necessary information to make a social choice is either the top ranked alternatives or the bottom ranked alternatives by the agents. If the rule relies on information on the top ranked alternatives, then, the alternatives chosen by the plurality rule are contained in the value of the rule. If the rule relies on information on the bottom ranked alternatives, then the alternatives chosen by the antiplurality rule are contained in the value of the rule.

Because the antiplurality rule is not efficient[6] when $m \geq 3$ and it is equal to the plurality rule when $m = 2$, Theorem 3.2 readily implies the following theorem.

Theorem 3.3 *If a rule $(\mathcal{M}_N, f)$ operates on minimal informational requirements in $\mathcal{ANMP}$, then it is a supercorrespondence of the plurality rule.*

This theorem gives a new characterization of the plurality rule; it is the most selective rule among the rules operating on minimal informational requirements in $\mathcal{ANMP}$. When you want to choose the socially best alternatives, then it is natural to adopt a rule in $\mathcal{ANMP}$. Theorem 3.3 shows that if you care for the informational processing cost and selectivity, then the answer is the plurality rule.

We conclude this section with the following remark. We defined the informational size of $(\mathcal{M}_N, f)$ simply by $\sum_{i \in N} |\mathcal{M}_i|$. Our results do not depend on this specific way of defining the informational size. Let g be any strictly increasing function on the positive orthant of $\mathbb{R}^n$, the n-dimensional Euclidean space, and let us define $g(|\mathcal{M}_1|, \ldots, |\mathcal{M}_n|)$ to be the informational size of $(\mathcal{M}_N, f)$. Then, we can obtain the same results with this definition of the informational size.

4 Proofs

In this section, we introduce many permutations of N and X. For simplicity, when we describe a permutation, we do not specify the part on which the permutation is the identity function. For example, when we say that σ is the permutation of N exchanging i and j, then it should be understood that σ is the identity function on $N \setminus \{i, j\}$.

[6]For example, let $X = \{x, y, z\}$ and let R_N be a preference profile such that xR_iyR_iz for all $i \in N$. Then, the antiplurality rule chooses $\{x, y\}$ while y is dominated by x.

4.1 Proof of Theorem 3.1

We proceed to establish Theorem 3.1 through a series of lemmas. Let $(\mathcal{M}_N, f)$ be a rule which operates on minimal informational requirements in $\mathcal{AN}$. Because the plurality rule is in $\mathcal{AN}$, the informational size of $(\mathcal{M}_N, f)$ is not greater than nm.

Lemma 4.1 *$\mathcal{M}_i = \mathcal{M}_j$ for all $i, j \in N$.*

Proof. Suppose to the contrary that $\mathcal{M}_i \neq \mathcal{M}_j$ for some $i, j \in N$.
CLAIM 1: At least one of the following two statements holds:

(i) There exist $M_i^* \in \mathcal{M}_i$ and $M_j^1, M_j^2 \in \mathcal{M}_j$ such that $M_i^* \cap M_j^1 \neq \emptyset$ and $M_i^* \cap M_j^2 \neq \emptyset$.

(ii) There exist $M_i^1, M_i^2 \in \mathcal{M}_i$ and $M_j^* \in \mathcal{M}_j$ such that $M_i^1 \cap M_j^* \neq \emptyset$ and $M_i^2 \cap M_j^* \neq \emptyset$.

Proof of Claim 1. Suppose that neither of the statements holds. Because (i) does not hold, for any $M_i \in \mathcal{M}_i$, there exists M_j such that $M_i \subset M_j$. Because (ii) does not hold, for any $M_j \in \mathcal{M}_j$, there exists $M_i \in \mathcal{M}_i$ such that $M_j \subset M_i$. Thus, for any $M_i \in \mathcal{M}_i$, there exist $M_j \in \mathcal{M}_j$ and $M_i' \in \mathcal{M}_i$ such that $M_i \subset M_j \subset M_i'$. Because $\mathcal{M}_i$ is a partition of $\mathcal{L}$, this implies that $M_i = M_j$. Therefore, $\mathcal{M}_i = \mathcal{M}_j$, which is a contradiction. □

Without loss of generality, assume that statement (i) of Claim 1 holds.
CLAIM 2: $f(M_j^1, M_{-j}) = f(M_j^2, M_{-j})$ for all $M_{-j} \in \mathcal{M}_{-j}$.
Proof of Claim 2. Suppose to the contrary that $f(M_j^1, M_{-j}) \neq f(M_j^2, M_{-j})$ for some $M_{-j} \in \mathcal{M}_{-j}$. Let R_j and R_j' be such that $R_j \in M_i^* \cap M_j^1$ and $R_j' \in M_i^* \cap M_j^2$. Let R_{-j} be an element of M_{-j}. $f(\varphi_N(R_j, R_{-j})) \neq f(\varphi_N(R_j', R_{-j}))$. Now, interchange the preferences of agents i and j. (Let σ denote the permutation interchanging agents i and j.) Then, by anonymity, $f(\varphi_N([(R_j, R_{-j})^\sigma])) \neq f(\varphi_N([(R_j', R_{-j})^\sigma]))$. However, because $R_j, R_j' \in M_i^*$, $\varphi_N([(R_j, R_{-j})^\sigma]) = \varphi_N([(R_j', R_{-j})^\sigma])$, which is a contradiction. □

Claim 2 implies that distinct messages M_j^1 and M_j^2 can be integrated into one message without any essential change. Formally, let $\mathcal{M}_j' = \{M_j \mid M_j \in \mathcal{M}_j \setminus \{M_j^1, M_j^2\}$ or $M_j = M_j^1 \cup M_j^2\}$. For $i \in N \setminus \{j\}$, let $\mathcal{M}_i' = \mathcal{M}_i$. Let $\mathcal{M}_N' = \prod_{i \in N} \mathcal{M}_i'$. For each message profile $M_N \in \mathcal{M}_N'$, let $f'(M_N) = f(M_N)$ if $M_j \neq M_j^1 \cup M_j^2$ and $f'(M_N) = f(M_j^1, M_{-j})$ if $M_j = M_j^1 \cup M_j^2$. We claim that $f'(\varphi_N'(R_N)) = f(\varphi_N(R_N))$ for every preference profile R_N. If $R_j \notin M_j^1 \cup M_j^2$, then $f'(\varphi_N'(R_N)) = f'(M_N) = f(M_N) = f(\varphi_N(R_N))$. If $R_j \in M_j^1$, then $f'(\varphi_N'(R_N)) = f'(M_j^1 \cup M_j^2, M_{-j}) = f(M_j^1, M_{-j}) = f(\varphi_N(R_N))$. If $R_j \in M_j^2$, then $f'(\varphi_N'(R_N)) = f'(M_j^1 \cup M_j^2, M_{-j}) = f(M_j^1, M_{-j}) = f(M_j^2, M_{-j}) = f(\varphi_N(R_N))$. ($\varphi_N'$ is a profile of message functions associated with $\mathcal{M}_N'$.) Therefore, $(\mathcal{M}_N', f')$ is in $\mathcal{AN}$ whereas the informational size of $(\mathcal{M}_N', f')$ is less than that of $(f, \mathcal{M}_N)$, which is a contradiction to the fact that $(\mathcal{M}_N, f)$ attains the minimal informational size in $\mathcal{AN}$. ■

Consider the case $m = 2$. Let $X = \{x, y\}$, let R_i be the linear order such that $r_1(R_i) = x$ and $r_2(R_i) = y$ and let R_i' be the linear order such that $r_1(R_i') = y$ and $r_2(R_i') = x$. Then, by Lemma 4.1, either $\mathcal{M}_i = \{\{R_i, R_i'\}\}$ for all $i \in N$ or $\mathcal{M}_i = \{\{R_i\}, \{R_i'\}\}$ for all $i \in N$. In the former case holds, because there is only one possible message profile, $f \circ \varphi_N$ should be constant on $\mathcal{L}^N$, which is a contradiction. Thus, the latter case holds. Therefore, we complete the proof of Theorem 3.1 for the case $m = 2$. (h can be either 1 or 2.) In the following, we assume $m \geq 3$.

Lemma 4.2 *For any $i \in N$, for any permutation ρ of X, and for any $M \in \mathcal{M}_i$, $\rho(M) \in \mathcal{M}_i$.*

Proof. Suppose $\rho(M) \notin \mathcal{M}_i$ for some $M \in \mathcal{M}_i$. There are two cases to consider.
CASE 1: $\rho(M) \subsetneq M'$ for some $M' \in \mathcal{M}_i$. Because $M \subsetneq \rho^{-1}(M')$, there exists $M^* \in \mathcal{M}_i$ such that $M^* \neq M$ and $M^* \cap \rho^{-1}(M') \neq \emptyset$.
CLAIM: $f(M, M_{-i}) = f(M^*, M_{-i})$ for all $M_{-i} \in \mathcal{M}_{-i}$.

Proof of Claim. Suppose to the contrary that $f(M, M_{-i}) \neq f(M^*, M_{-i})$ for some $M_{-i} \in \mathcal{M}_{-i}$. Let R_i be any element of M, let R_{-i} be any element of M_{-i}, and let $\hat{R}_i$ be any element of $\rho^{-1}(M') \cap M^*$. Then, $(f \circ \varphi_N)(R_i, R_{-i}) \neq (f \circ \varphi_N)(\hat{R}_i, R_{-i})$. By neutrality, $(f \circ \varphi_N)[\rho(R_i), \rho(R_{-i})] \neq (f \circ \varphi_N)[\rho(\hat{R}_i), \rho(R_{-i})]$. However, because $\rho(R_i), \rho(\hat{R}_i) \in M'$, $\varphi_N[\rho(R_i), \rho(R_{-i})] = \varphi_N[\rho(\hat{R}_i), \rho(R_{-i})]$, which is a contradiction. □

This claim shows that we can integrate distinct messages M and M^* into one message without any substantial change. See the argument following Claim 2 in the proof of Lemma 4.1. The same reasoning applies here, and we have a contradiction.

CASE 2: $\rho(M) \cap M^1 \neq \emptyset$ and $\rho(M) \cap M^2 \neq \emptyset$ for some $M^1, M^2 \in \mathcal{M}_i$. In this case, we claim $f(M^1, M_{-i}) = f(M^2, M_{-i})$ for all $M_{-i} \in \mathcal{M}_{-i}$. Suppose not. Then, for any $R_i^1 \in \rho(M) \cap M^1$ and for any $R_i^2 \in \rho(M) \cap M^2$, $(f \circ \varphi_N)(R_i^1, R_{-i}) \neq (f \circ \varphi_N)(R_i^2, R_{-i})$. By neutrality, $(f \circ \varphi_N)(\rho^{-1}(R_i^1), \rho^{-1}(R_{-i})) \neq (f \circ \varphi_N)(\rho^{-1}(R_i^2), \rho^{-1}(R_{-i}))$. However, because $\rho^{-1}(R_i^1), \rho^{-1}(R_i^2) \in M$, we have $\varphi_N(\rho^{-1}(R_i^1), \rho^{-1}(R_{-i})) = \varphi_N(\rho^{-1}(R_i^2), \rho^{-1}(R_{-i}))$, which is a contradiction.

Thus, $f(M^1, M_{-i}) = f(M^2, M_{-i})$ for all $M_{-i} \in \mathcal{M}_{-i}$. This implies that we can integrate M^1 and M^2 into one message without affecting any essential aspects of a rule $(\mathcal{M}_N, f)$. By the same argument as in the proof of Lemma 4.1, we have a contradiction. ■

Lemma 4.3 *For any $i \in N$, there exists $h \in \{1, \ldots, m\}$ such that for any $M \in \mathcal{M}_i$, $r_h(M) = \{x \in X \mid r_h(R_i) = x$ for some $R_i \in M\}$ is a singleton.*

Proof. Suppose to the contrary that for any $h \in \{1, \ldots, m\}$, there exists $M \in \mathcal{M}_i$ such that $r_h(M)$ is not a singleton. Let M' be any element of $\mathcal{M}_i$ and let R_i and R_i' be any elements of M and M', respectively. Let ρ be the permutation of X such that $\rho(R_i) = R_i'$. Then, $\rho(M) \cap M' \neq \emptyset$. By Lemma 4.2, $\rho(M) \in \mathcal{M}_i$. Because $\mathcal{M}_i$ is a partition of $\mathcal{L}$, $\rho(M) = M'$. This implies that $r_h(M')$ is not a singleton. This argument shows that for any $h \in \{1, \ldots, m\}$ and for any $M \in \mathcal{M}_i$, there exist $R, R' \in M$ such that $r_h(R) \neq r_h(R')$.

CLAIM 1: For any $h \in \{1, \ldots, m\}$, for any $M \in \mathcal{M}_i$, and for any $x \in X$, there exists $R \in M$ such that $r_h(R) = x$. In other words, $r_h(M) = X$ for all $h \in \{1, \ldots, m\}$ and $M \in \mathcal{M}_i$.

Proof of Claim 1. Suppose not. Then, there exist $h \in \{1, \ldots, m\}$ and $M \in \mathcal{M}_i$ such that $r_h(M) \neq X$. We claim that $|r_h(M)| = m - 1$.

Suppose $|r_h(M)| \leq m - 2$. Let $X \setminus r_h(M) = \{y_1, \ldots, y_{h_1}\}$ and let $r_h(M) = \{x_1, \ldots, x_{h_2}\}$. Because $|r_h(M)| \leq m - 2$, $h_1 \geq 2$. Because $r_h(M)$ is not a singleton, $h_2 \geq 2$. For each pair (ℓ_1, ℓ_2) such that $1 \leq \ell_1 \leq h_1$ and $1 \leq \ell_2 \leq h_2$, let $\rho_{\ell_1}^{\ell_2}$ be the permutation exchanging y_{ℓ_1} and x_{ℓ_2}. Then, $M \neq \rho_{\ell_1}^{\ell_2}(M) \neq \rho_{\ell_1'}^{\ell_2'}(M)$ for any $\ell_1, \ell_1', \ell_2, \ell_2'$ with $(\ell_1, \ell_2) \neq (\ell_1', \ell_2')$. By Lemma 4.2, $\rho_{\ell_1}^{\ell_2}(M) \in \mathcal{M}_i$ for all ℓ_1, ℓ_2. Thus, $|\mathcal{M}_i| \geq h_1 \cdot h_2 + 1 \geq 2 \cdot \max\{h_1, h_2\} + 1 \geq m + 1 > m$. Then, by Lemma 4.1, the informational size of $(\mathcal{M}_N, f)$ is greater than nm, which is a contradiction. Thus, $|r_h(M)| = m - 1$.

Let $\{x\} = X \setminus r_h(M)$. Let R be any element of M and let h' be such that $r_{h'}(R) = x$. Because $r_{h'}(M)$ is not a singleton (see the statement right above Claim 1), there exists $R' \in M$ such that $r_{h'}(R') \neq x$. Let h'' be such that $r_{h''}(R') = x$. Note that $h' \neq h''$ and $h', h'' \neq h$. Let y denote $r_{h''}(R)$ and let ρ be the permutation of X such that $\rho(R) = R'$. Then, because $\rho(M) \cap M \neq \emptyset$ and $\mathcal{M}_i$ is a partition of $\mathcal{L}$, $\rho(M) = M$. Note that $\rho(y) = x$. Because $y \in r_h(M)$, there exists $R'' \in M$ such that $r_h(R'') = y$. For such R'', $r_h(\rho(R'')) = x$, which is a contradiction to $\rho(R'') \in M$. □

CLAIM 2: $f(M_N) = X$ for all $M_N \in \mathcal{M}_N$.

Proof of Claim 2. Suppose to the contrary that $f(M_N) \neq X$ for some $M_N \in \mathcal{M}_N$. Let R_N be any element of M_N. Then, $(f \circ \varphi_N)(R_N) \neq X$. Let x be an element of $X \setminus (f \circ \varphi_N)(R_N)$. By neutrality, there exists R_N' such that $x \in (f \circ \varphi_N)(R_N')$. Let M_N' be the element of $\mathcal{M}_N$ such that $R_N' \in M_N'$.

$r_h(R'_i)$	$r_h(R_i)$	Operation on R_i	Operation on R''_i
x	x	Do not change	Do not change
	Not x	Interchange $r_{h+1}(R_i)$ and x	Lift x to hth position
Not x (Let $y = r_h(R'_i)$)	y	Do not change	Do not change
	x	Interchange $r_{h-1}(R_i)$ and y	Lift x to the top
	Others	First, interchange x and $r_m(R_i)$ and next, interchange y and $r_{h-1}(R_i)$	Lift x to the top

Table 1: The profiles R''_N and R^*_N in the proof of Theorem 3.2

Let i be any agent and let h be such that $r_h(R_i) = x$. By Claim 1, $r_h(M'_i) = X$. Thus, in M'_i, we can find R''_i such that $r_h(R''_i) = x$. Let R''_N be a profile of such R''_i. Note that the positions of x in R''_N are the same as in R_N. Also, because R''_N belongs to M'_N, $\varphi'_N(R'_N) = \varphi'_N(R''_N)$. Thus, $x \in (f \circ \varphi_N)(R''_N)$. Let ρ_N be a profile of permutations such that $\rho_i(R''_i) = R_i$. Note that $\rho_i(x) = x$ for all $i \in N$. By neutrality, $x \in (f \circ \varphi_N)(\rho(R''_N)) = (f \circ \varphi_N)(R_N)$, which is a contradiction. □

Claim 2 implies that a rule (f, φ_N) is constant, which is a contradiction. ■

Lemma 4.3 shows that each $M_i \in \mathcal{M}_i$ is contained in $\{R_i \in \mathcal{L} \mid r_h(R_i) = x\}$ for some $x \in X$. Thus, if the hth ranked alternatives in two preferences R and R' are different, then R and R' belong to distinct M_i and M'_i in $\mathcal{M}_i$. This implies that there are at least m elements in $\mathcal{M}_i$. If $|\mathcal{M}_i| > m$, then by Lemma 4.1, the informational size of $(\mathcal{M}_N, f)$ is greater than nm, which is a contradiction. Thus, $|\mathcal{M}_i| = m$ for all $i \in N$, and the informational size of $(\mathcal{M}_N, f)$ is nm.

For each $M_i \in \mathcal{M}_i$, let $C(M_i)$ denote the element of X such that $M_i \subset \{R_i \in \mathcal{L} \mid r_h(R_i) = C(M_i)\}$. We show that this C is a bijection. Because $|\mathcal{M}_i| = m = |X|$, it suffices to show that C is onto. Let x be any element of X. Then, because $\{R_i \in \mathcal{L} \mid r_h(R_i) = x\}$ is not the empty set and $\mathcal{M}_i$ is a partition of $\mathcal{L}$, there exists $M_i \in \mathcal{M}_i$ such that $M_i \subset \{R_i \in \mathcal{L} \mid r_h(R_i) = x\}$, and hence $C(M_i) = x$. Thus, C is a bijection. This implies that for any $M_i \in \mathcal{M}_i$, for any $M'_i \in \mathcal{M}_i \setminus \{M_i\}$, and for any $R_i \in M'_i$, $r_h(R_i)$ is not $C(M_i)$. Thus, $M_i \subsetneq \{R_i \in \mathcal{L} \mid r_h(R_i) = C(M_i)\}$ leads to a contradiction to the fact that $\mathcal{M}_i$ is a partition of $\mathcal{L}$. Therefore, for each $M_i \in \mathcal{M}_i$, $M_i = \{R_i \in \mathcal{L} \mid r_h(R_i) = C(M_i)\}$. That is, each $M_i \in \mathcal{M}_i$ is associated with an alternative $C(M_i)$ in X and M_i consists of all preferences which rank $C(M_i)$ at the hth position. Because C is a bijection, we complete the proof of the Theorem 3.1.

4.2 Proof of Theorem 3.2

Let $(\mathcal{M}_N, f)$ be a rule which operates on minimal informational requirements in $\mathcal{ANM}$.

First, we prove the statement (i). If $m = 2$, then this statement is a direct consequence of Theorem 3.1. Thus, let $m \geq 3$. Suppose to the contrary that $1 < h < m$, and we claim that $(\mathcal{M}_N, f)$ is constant. Let R_N and R'_N be any preference profiles. We prove $(f \circ \varphi_N)(R_N) = (f \circ \varphi_N)(R'_N)$.

First, we show $(f \circ \varphi_N)(R_N) \subset (f \circ \varphi_N)(R'_N)$. Let x be any element of $(f \circ \varphi_N)(R_N)$. Now, make a new preference profile R''_N from R_N according to the third column of Table 7.1. (At this stage, see only the first three columns.) Depending on $r_h(R'_i)$ and $r_h(R_i)$, there are five possible

cases as described in the first two columns of Table 7.1. The third column specifies the operation on R_i in each case. Note that these operations are feasible because neither $h = 1$ nor $h = m$.

Let R''_N denote the resulting preference profile. It can be seen that $r_h(R_i) = r_h(R''_i)$ for all $i \in N$. Thus, $(f \circ \varphi_N)(R_N) = (f \circ \varphi_N)(R''_N)$ and x is also in $(f \circ \varphi_N)(R''_N)$. Now, apply the operation on R''_N described in the forth column of Table 7.1, and let R^*_N denote the resulting preference profile. Then, by monotonicity, $x \in (f \circ \varphi_N)(R^*_N)$. It can be seen that $r_h(R'_i) = r_h(R^*_i)$ for all $i \in N$, and hence $(f \circ \varphi_N)(R'_N) = (f \circ \varphi_N)(R^*_N)$. Therefore, $x \in (f \circ \varphi_N)(R'_N)$, and we complete the proof of the relation $(f \circ \varphi_N)(R_N) \subset (f \circ \varphi_N)(R'_N)$.

By the symmetric argument, we can prove $(f \circ \varphi_N)(R_N) \supset (f \circ \varphi_N)(R'_N)$.

Because R_N and R'_N was arbitrary, we can conclude $(\mathcal{M}_N, f)$ is a constant rule, which is a contradiction. Thus, h should be either 1 or m.

Next, we prove the second statement of the theorem.

CASE 1: $h = 1$. By Theorem 3.1, $\mathcal{M}_N$ is equal to the domain of f^p. In this case, we prove $f^p(M_N) \subset f(M_N)$ for all $M_N \in \mathcal{M}_N$. Suppose $f^p(M_N) \not\subset f(M_N)$ for some $M_N \in \mathcal{M}_N$. Let x be an element of $f^p(M_N) \setminus f(M_N)$, and let R_N be such that $R_i \in M_i$ for all $i \in N$.

We claim that $f^p(M_N) \cap f(M_N) = \emptyset$. Suppose to the contrary that there exists $y \in f^p(M_N) \cap f(M_N)$. Then, $y \in (f^p \circ \varphi^p_N)(R_N) \cap (f \circ \varphi_N)(R_N)$. (Note that $\varphi^p_N = \varphi_N$.) Let σ be a permutation of N such that $\sigma(\{i \in N \mid r_1(R_i) = x\}) = \{i \in N \mid r_1(R_i) = y\}$ and $\sigma(\{i \in N \mid r_1(R_i) = y\}) = \{i \in N \mid r_1(R_i) = x\}$. By anonymity, $y \in (f \circ \varphi_N)(R^\sigma_N)$. Let ρ be the permutation of X exchanging x and y. By neutrality, $x \in (f \circ \varphi_N)(\rho(R^\sigma_N))$. Note that the two n-tuples of top ranked alternatives in R_N and $\rho(R^\sigma_N))$ are the same. Because $h = 1$, $(f \circ \varphi_N)(R_N) = (f \circ \varphi_N)(\rho(R^\sigma_N))$. Thus, $x \in (f \circ \varphi_N)(R_N) = f(M_N)$, which is a contradiction.

Let z be any element of $(f \circ \varphi_N)(R_N)$. By the above argument, $z \notin (f^p \circ \varphi^p_N)(R_N)$. Let N_x be a subset of $\{i \in N \mid r_1(R_i) = x\}$ such that $|N_x| = |\{i \in N \mid r_1(R_i) = z\}|$. Then, let σ' be the permutation such that $\sigma'(N_x) = \{i \in N \mid r_1(R_i) = z\}$ and $\sigma'(\{i \in N \mid r_1(R_i) = z\}) = N_x$. By anonymity, $z \in (f \circ \varphi_N)(R^{\sigma'}_N)$. Let ρ' be the permutation exchanging x and z. By neutrality, $x \in (f \circ \varphi_N)(\rho'(R^{\sigma'}_N))$. For each $i \in \{j \in N \mid r_1(R_j) = x\} \setminus N_x$, lift x to the top in $\rho'(R^{\sigma'}_i)$. Let R''_N denote the resulting preference profile. By monotonicity, $x \in (f \circ \varphi_N)(R''_N)$. It can be seen that the two n-tuples of top ranked alternatives in R_N and R''_N are the same, and hence $x \in (f \circ \varphi_N)(R_N) = f(M_N)$, which is a contradiction. Therefore, $f^p(M_N) \subset f(M_N)$ for all $M_N \in \mathcal{M}_N$.

CASE 2: $h = m$. Suppose $f^a(M_N) \not\subset f(M_N)$ for some $M_N \in \mathcal{M}_N$. Let x be an element of $f^a(M_N) \setminus f(M_N)$, and let R_N be such that $R_i \in M_i$ for all $i \in N$.

We claim that $f^a(M_N) \cap f(M_N) = \emptyset$. Suppose to the contrary that there exists $y \in f^a(M_N) \cap f(M_N)$. Then, $y \in (f^a \circ \varphi^a_N)(R_N) \cap (f \circ \varphi_N)(R_N)$. There are two cases to consider.

First, assume $\{i \in N \mid r_m(R_i) = y\} = \emptyset$. Then, $\{i \in N \mid r_m(R_i) = x\}$ is also the empty set. Let ρ be the permutation of X exchanging x and y. By neutrality, $x \in (f \circ \varphi_N)(\rho(R_N))$. Note that the two n-tuples of the bottom ranked alternatives in R_N and $\rho(R_N)$ are the same. Thus, $x \in (f \circ \varphi_N)(R_N) = f(M_N)$, which is a contradiction.

Next, assume $\{i \in N \mid r_m(R_i) = y\} \neq \emptyset$. Then, $|\{i \in N \mid r_m(R_i) = y\}| = |\{i \in N \mid r_m(R_i) = x\}| > 0$. Let σ be a permutation of N such that $\sigma(\{i \in N \mid r_m(R_i) = y\}) = \{i \in N \mid r_m(R_i) = x\}$ and $\sigma(\{i \in N \mid r_m(R_i) = x\}) = \{i \in N \mid r_m(R_i) = y\}$. By anonymity, $y \in (f \circ \varphi_N)(R^\sigma_N)$. Let ρ be the permutation of X exchanging x and y. By neutrality, $x \in (f \circ \varphi_N)(\rho(R^\sigma_N))$. Note that the two n-tuples of the bottom ranked alternatives in R_N and $\rho(R^\sigma_N)$ are the same. Thus, $x \in (f \circ \varphi_N)(R_N)$, which is a contradiction. Therefore, in any case, $f^a(M_N) \cap f(M_N) = \emptyset$.

Let z be any element of $(f \circ \varphi_N)(R_N)$. By the above argument, $z \notin (f^a \circ \varphi^a_N)(R_N)$, that is, $|\{i \in N \mid r_m(R_i) = z\}| > |\{i \in N \mid r_m(R_i) = x\}|$. Let N_z be a subset of $\{i \in N \mid r_m(R_i) = z\}$ such that $|N_z| = |\{i \in N \mid r_m(R_i) = x\}|$. Let σ' be a permutation of N such that $\sigma'(N_z) = \{i \in$

$N \mid r_m(R_i) = x\}$ and $\sigma'(\{i \in N \mid r_m(R_i) = x\}) = N_z$. By anonymity, $z \in (f \circ \varphi_N)(R_N^{\sigma'})$. Let ρ' be the permutation of X exchanging x and z. By neutrality, $x \in (f \circ \varphi_N)(\rho(R_N^{\sigma'}))$. Now, for each $i \in \{j \in N \mid r_m(R_j) = z\} \setminus N_z$, take z to the second place from the bottom at $\rho(R_i^{\sigma'})$. Let R'_N be the resulting preference profile. Note that the two n-tuples of bottom ranked alternatives in $\rho(R_N^{\sigma})$ and R'_N are the same, and hence $x \in (f \circ \varphi_N)(R'_N)$. Now, for each $i \in \{j \in N \mid r_m(R_j) = z\} \setminus N_z$, lift x to the top of his preference. Let R''_N denote resulting preference profile. By monotonicity, $x \in (f \circ \varphi_N)(R''_N)$. Then, it can be seen that the two n-tuples of bottom ranked alternatives in R''_N and R_N are the same. Thus, $x \in (f \circ \varphi_N)(R_N)$, which is a contradiction. Therefore, $f^a(M_N) \subset f(M_N)$ for all $M_N \in \mathcal{M}_N$.

References

Ching, Stephen (1996) A simple characterization of plurality rule. Journal of Economic Theory 71, 298–302.

Conitzer, Vincent; Sandholm, Tuomas (2005) Communication complexity of common voting rules. Proceedings of the ACM Conference on Electronic Commerce, 78–87.

Kushilevitz, Eyal; Nisan, Noam (1997) Communication complexity. Cambridge University Press, Cambridge.

Moulin, Hervé (1980) On strategy-proofness and single peakedness. Public Choice 35, 437–455.

Moulin, Hervé (1988) Axioms of cooperative decision making. Cambridge University Press, Cambridge.

Richelson, Jeffrey T (1978) A characterization result for the plurality rule. Journal of Economic Theory 19, 548–550.

Roberts, Fred S. (1991) Characterization of the plurality function. Mathematical Social Sciences 21, 101–127.

Sen, Amartya K. (1986) Social choice theory. In: Arrow, Kenneth J.; Intrigator, Michael D. (eds) Handbook of mathematical economics, vol 3. North-Holland, Amsterdam, pp 1073–1181.

Yao, Andrew Chi-Chih (1979) Some complexity questions related to distributed computing. Proceedings of the 11th ACM symposium on theory of computing, 209–213.

Yeh, Chun-Hsien (2008) An efficiency characterization of plurality rule in collective choice problems. Economic Theory 34, 575–583.

Shin Sato
Graduate School of Economics, Keio University,
2-15-45 Mita, Minatoku, Tokyo, Japan
Email: sato@gs.econ.keio.ac.jp

Non-dictatorial Social Choice Rules Are Safely Manipulable

Arkadii Slinko and Shaun White

Abstract

When a number of like-minded voters vote strategically and have limited abilities to communicate the under and overshooting may occur when too few or too many of them vote insincerely. In this paper we discuss this phenomenon and define the concept of a safe strategic vote. We prove that for any onto and non-dictatorial social choice rule there exist a profile at which a voter can make a safe strategic vote. This means that on occasion a voter will have an incentive to make a strategic vote and know that he will not be worse off regardless of how other voters with similar preferences would vote, sincerely or not. We also extend the Gibbard-Satterthwaite theorem. We prove that an onto, non-dictatorial social choice rule which is employed to choose one of at least three alternatives is safely manipulable by a single voter. We discuss new problems related to computational complexity that appear in this new framework.

1 Introduction

The classical Gibbard-Satterthwaite theorem (1973–75) claims that for any non-dictatorial social choice rule which has at least three alternatives in its range, a voter, on occasion, will have an opportunity to vote strategically. This opportunity allows her to change the result and get a better outcome than the one that she would get if voted sincerely, ceteris paribus. However such occasions become rare when the number of voters is large. This statement, known as Pattanaik's conjecture ([13], p.102) was made rigorous — see, for example, [15, 16, 17, 18], — where, in particular, it was proved that under the Impartial Culture (IC) assumption, the probability of obtaining a manipulable profile is bounded from above by a scalar multiple of $1/\sqrt{n}$, where n is the number of voters. So, in large scale elections individual manipulability is not really an issue, as Pattanaik suspected.

However the issue of group manipulability, which is sometimes called coalitional manipulability, remains an issue since the probability of having a group of voters that can successfully manipulate does not go to zero when n grows [11]. It was shown [17] that, under the IC, to be able to manipulate with non-zero probability, a group of voters must include a scalar multiple of $\sqrt{n}$ voters. Moreover, in [12] it was shown that the average size of minimal manipulating coalition also contain a scalar multiple of $\sqrt{n}$ voters. Thus any group that wants to have a chance to manipulate must be large.

In practice there are several significant barriers for group manipulation happening. Firstly, this manipulating coalition must be somehow formed. If it has to be formed endogenously its formation, given its size, must be complex with a lot of private communication.

Secondly, even if it is exogenously defined, this group must include a coordinator who calculates who should submit which linear order and then privately communicates those to coalition members. Thirdly, all the coalition members must obey the instructions of the coordinator and there does not seem to be obvious ways to reinforce the discipline. In any case it is extremely unlikely that such a coalition can be formed in the absence of means of private voter-to-voter communication.

Here we would like to suggest a new model for a group formation of a manipulating coalition. We assume that it is possible for a voter to send a single message to the whole electorate (say through the media) but it is not possible to send a large number of "individualised" messages. The example of such a communication would be an important public figure calling upon her supporters to vote in a certain way. Obviously that only voters who share their views with this public figure might respond and change their vote accordingly. However, not all of them will respond; some of them might consider voting strategically unethical.

We make the classical social choice assumption that voters know sincere preferences of others but do not know their voting intentions. Therefore issuing a call to supporters the public figure will not know exactly how many supporters will follow her example and vote as she recommends. If the value of the social choice function may not drop below the status quo, then we say that such call is safe.

Due to the uncertainty in the number of like-minded voters attempting to vote strategically, sometimes an overshooting may occur when too many like-minded voters act strategically, and as a result the value of the social choice function drops below of what it would be if everybody voted sincerely. The same thing may happen if too few like-minded voters act on the incentive to manipulate; in such a case we talk about undershooting. If one of these situations (or both) is possible we classify such a strategic vote as unsafe. Making a safe strategic vote is safe in the sense that the value of the social choice function would not drop below the status quo no matter how many like-minded voter would respond and join in the manipulation attempt.

The main result of this paper states that if there are at least three alternatives, then for any non-dictatorial and onto social choice function a profile can be found at which one (or more) voter can make a safe strategic vote.

The issue of overshooting can also appear in the context of the Gibbard-Satterthwaithe theorem. Suppose that we have a profile which is manipulable by a single voter. Then it will be most likely manipulable by many voters (at least if the function is anonymous, then any voter with the same preferences will have the same incentive to manipulate). It might be the case that the value of the social choice function gets initially better, when one voter votes insincerely but then gradually drops even below the status quo as more and more like-minded voters join in a manipulating attempt. We call such individual manipulation unsafe and possible overshooting will be a significant deterrent to such manipulation.

Unfortunately such deterrent does not always exist. As a corollary to our main result we extend the Gibbard-Satterthwaithe theorem by showing that any onto non-dictatorial social

choice rule is, on occasion, safely manipulable if more than three alternatives are involved.

There is an extensive research into the computational complexity of strategic voting. Computational complexity of manipulating classical voting procedures has been studied, for example, in Bartholdi *et al* [2], Bartholdi and Orlin [3] and, more recently, Conitzer and Sandholm [4, 6], Faliszewski [9] and Conitzer [7]. Some artificial protocols for which manipulation is hard (in worst case scenario) have been also designed by Conitzer and Sandholm [5] and Elkind and Lipmaa [8]. At the same time it has been shown that no voting protocol is hard to manipulate in the average case scenario [6]. It would be interesting to extend the complexity analysis to the new type of manipulation introduced in this paper.

To the best of our knowledge, the distinction between safe and unsafe strategic votes, as we define them, was first made (albeit in the context of parliament choosing rules) in Slinko and White [19].

2 Strategic overshooting and undershooting

For the remainder of this section let us fix the set of alternatives $\mathcal{A}$, a set of voters $[n] = \{1, 2, \ldots, n\}$, and a social choice rule F. Preferences of each voter i are represented by a linear order R_i on $\mathcal{A}$ and the sequence $R = (R_1, \ldots, R_n)$ is called a *profile*. We will say that two voters i and j are of the *same type* if they have identical preferences, i.e. $R_i = R_j$. The type of the voter i is denoted as $<i>$. It is identified with R_i. Voters of the same type will be also called *like-minded*.

As usual, if $V \subseteq [n]$, by $R_{-V}(L)$ we will denote the profile obtained from R when all voters from V vote L and all other voters retain their original linear orders. For $V \subseteq [n]$ we will write $a \succ_V b$ if all voters from V strictly prefer a to b.

Definition 1 (An incentive to vote strategically)**.** *Fix a voter i, and define V to be the set of all voters with preferences identical to those of i at R. If there exists a linear order $L \neq R_i$ over $\mathcal{A}$, and a subset $V_1 \subseteq V$ containing i such that*

$$F(R_{-V_1}(L)) \succ_V F(R)$$

then we will say that, at R, voter i has an incentive to vote strategically L.

This is the key concept of the paper. We note that to have incentive to vote strategically does not mean that the voter is pivotal. What this voter can hope for is that there will be a sufficient number of like-minded voters with preferences identical to hers who will make a strategic move. She can call upon them to do so.

In some circumstances (we will present an example in the next section) a voter may hesitate to act on an incentive to vote strategically or to issue a call. One reason for hesitation would be this: in attempting to vote strategically, the voter could realise a gain or could realise a loss depending on which other voters with the same preferences also vote strategically. We now describe such circumstances formally.

Definition 2 (Strategic overshooting)**.** *Fix a voter i, and define V to be the set of all voters of type $<i>$ at R. Suppose that there exist two sets V_1 and V_2 such that $i \in V_1 \subset V_2 \subseteq V$, and a linear order $L \neq R_i$ such that:*

- *every voter in V_2 has an incentive to strategically vote L, and*
- *$F(R_{-V_1}(L)) \succ_V F(R) \succ_V F(R_{-V_2}(L))$.*

Then voter i (together with other voters of type $<i>$) can strategically overshoot *at R voting L.*

If we reverse the roles of the sets V_1 and V_2 we will obtain a definition of *strategic undershooting*. These two concepts are not mutually exclusive. Theoretically it is possible that a voter can both strategically overshoot and strategically undershoot at R with a vote of L.

Definition 3 (Safe and unsafe strategic votes)**.** *Fix a voter i, and a profile R. Suppose that there exists a linear order $L \neq R_i$ such that*

- *at R, voter i has an incentive to strategically vote L; and*
- *voter i cannot strategically overshoot or strategically undershoot at R with a vote of L.*

Then voter i can make a safe strategic vote *at R. If the first condition is satisfied but the second is not, we will say that voter i can make an* unsafe strategic vote *at R.*

It is important to note that the fact that the fact that a voter can make a safe strategic vote does not mean that she is pivotal. But, even if she is not pivotal, she has a strong incentive to vote strategically and try to rally supporters to vote likewise. If a strategic vote is unsafe, a voter has an incentive to vote strategically but she also has a disincentive, namely the prospect of making the outcome worse rather than better. In a similar circumstances, with respect to parliament choosing rules, Slinko and White [19] suggested that in this case voters will act in accord with their attitude towards uncertainty.

The main theorem of this paper is:

Theorem 1 (Slinko & White, 2008)**.** *Suppose an onto, non-dictatorial social choice rule is employed to choose one of at least three alternatives. Then there exist a profile at which a voter can make a safe strategic vote.*

Proof. The proof is rather technical and can be found in a preprint [20]. □

It has proved useful to classify a certain kind of strategic moves as an escape. Suppose preferences are such that voter i ranks no element of $\mathcal{A}$ lower than X. Further suppose that R is the profile of sincere preferences (over $\mathcal{A}$), and $F(R) = X$. If, at R, voter i has an incentive to vote strategically then voter i (together with other voters of type $<i>$) will be said to be able to *escape* at R. Notice that an escape is more than an ordinary safe strategic vote. Indeed, voter i cannot get worse no matter how all other voters vote (not only voters of type $<i>$). The concept of escaping appears often during the proofs.

3 Examples and Their Geometric Interpretation

Given any scoring social choice rule other than plurality, and a set of voters which is sufficiently large (more than 50, say), it is easy to create examples of strategic overshooting. In this section we will deal only with anonymous rules so we will use the so-called "succinct input" [9] which in Social Choice is known as voting situation [1]. Passing from a profile to a voting situation we forget the order of linear orders in the profile which makes votes anonymous. This is especially convenient when we have only three alternatives A, B, and C. In particular, the table

Preference order	Number of voters
ABC	n_1
ACB	n_2
BAC	n_3
BCA	n_4
CAB	n_5
CBA	n_6

denotes the voting situation when n_1 voters prefer A to B to C, n_2 voters prefer A to C to B, etc. This voting situation is denoted as $(n_1, n_2, n_3, n_4, n_5, n_6)$.

Example 1 (Strategic overshooting, escaping under the Borda rule)**.** *Suppose 94 voters are using the Borda rule to choose one of three alternatives and that the corresponding voting situation of sincere preferences is* $(17, 15, 18, 16, 14, 14)$*. If all voters vote sincerely then A would score 96, B 99, and C 87; B would win. If between four and eight ABC types vote ACB, ceteris paribus, then A would win. If 10 or more ABC types vote ACB, ceteris paribus, then C would win. So the given voting situation of sincere preferences is prone to unsafe manipulation. This voting situation is also prone to safe manipulation: if 13 or more ACB voters vote CAB, ceteris paribus, then C will win, and the manipulators will have made an escape.*

To explain Example 1 geometrically we need a graphical representation of the scores and strategic moves. Firstly we normalise the scores. Given weights $w_1 \geq w_2 \geq \ldots \geq w_m = 0$ we define the corresponding score function $sc\colon A \to \mathbb{R}$. For a profile $R = (R_1, \ldots, R_n)$ we set

$$sc(a) = \sum_{i=1}^{n} w_{pos(a,R_i)},$$

where $pos(a, R_i)$ is the position of a in R_i, i.e. the number of alternatives which are no worse than a relative to R_i. Then the *normalised positional score* of a candidate a is given by:

$$scn(a) = \frac{sc(a)}{sc(a_1) + \ldots + sc(a_m)}.$$

After this normalisation we have

$$scn(a_1) + scn(a_2) + \ldots + scn(a_m) = 1.$$

A normalised vector of scores $scn(a)$ can be represented as a point X of the m-dimensional simplex S^{m-1}:

$$\mathbf{x} = (x_1, \ldots, x_m), \qquad x_1 + \ldots + x_m = 1,$$

where $x_i = scn(a_i)$ is the normalised score of the ith alternative. We treat $x_1, \ldots, x_n$ as the homogeneous barycentric coordinates of X.

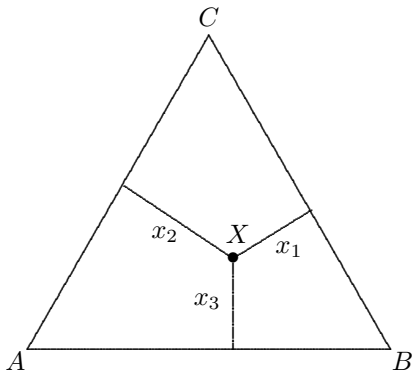


Figure 1: Geometric representation of a normalised score

Irrespective of the positional scoring rule, by voting insincerely a voter who prefers A to B to C cannot improve the score of A, nor can she worsen the score of C. If she votes insincerely, she will expect the vector of scores to fall in the shaded area.

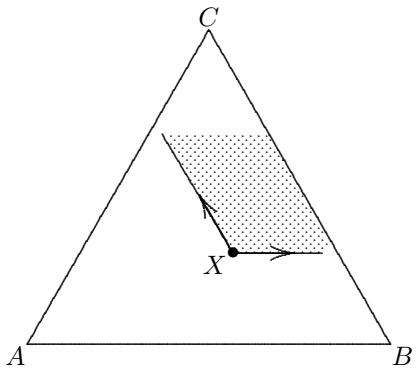


Figure 2: Possible directions of change under a manipulation attempt

By insincerely reporting her preferences to be BAC, she will move the vector of scores horizontally east. This she can do so long as the score function is not antiplurality. By

insincerely reporting ACB, she moves the vector of scores north west, parallel to BC, and this is possible except in the event the score function is plurality.

Now the geometry of Example 1 can be represented as follows: The simplex S^2 is split into four quadrilaterals $AKOM$, $BMOL$, $CLOK$, where the winners will be A, B and C, respectively. For the sincere profile we are in the quadrilateral $BMOL$ where B wins. When ABC types report ACB they move the vector of normalised scores in north-western direction parallel to BC. Overshooting occurs, when the normalised vector of scores crosses the quadrilateral $AKOM$, where A wins and penetrates $CLOK$, where the winner is C. If ACB voters vote CAB they move the vector of normalised scores from $BMOL$ to $CLOK$ without crossing $AKOM$. This precludes over and undershooting.

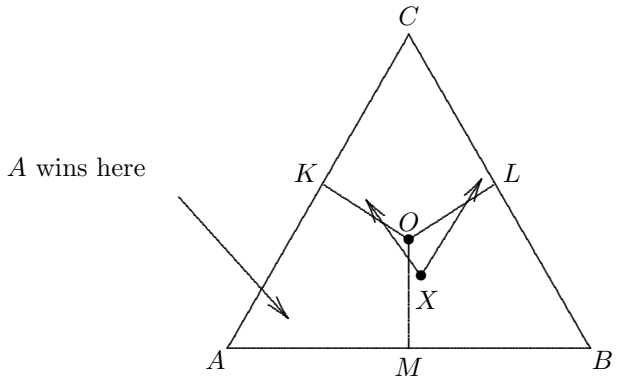


Figure 3: Geometric representation of the two moves from Example 1

The arrow parallel to BC graphically represents the unsafe manipulation that overshoots; it crosses the area where A wins and ends up in the area where the winner is C. The arrow parallel to AC represents the safe strategic move that is successful.

Here is yet another example for a multistage elimination rule..

Example 2 (Strategic overshooting under plurality with a run-off)**.** *Suppose 23 voters are using a plurality with a run-off social choice rule to choose one of three alternatives and suppose the corresponding voting situation is* $(4, 6, 7, 0, 0, 6)$*. If all voters vote sincerely then B beats A 13-10 in the run-off. If 2 voters of type ABC vote for C in the first round, ceteris paribus, then A beats C 17-6 in the run-off. If 4 or more voters of type ABC vote for C in the first round, ceteris paribus, then C beats B 12-11 in the run-off. Thus the profile of sincere preferences described is unsafely manipulable.*

We continue our series of examples with an example of a profile where only unsafe strategic vote can be made.

Example 3 (A profile that is unsafely but not safely manipulable)**.** *Suppose 80 voters are using a scoring social choice rule to choose one of three alternatives. Suppose that a first*

place ranking on a ballot is worth three points, while a second place ranking is worth one point, i.e. $w_1 = 3$, $w_2 = 1$, $w_3 = 0$. *Let the sincere voting situation be* $(30, 0, 20, 0, 0, 30)$. *If all voters are honest then A scores 110, B 120, and C 90; B wins. Those that rank B or C highest have no incentive to act strategically. Consider the voters of type ABC. If between 11 and 19 of them state they are of type ACB, ceteris paribus, then A will win. If more than 21 of them state they are of type ACB, ceteris paribus, then C will win. Voters of type ABC have no other way to manipulate the vote. Thus the profile of sincere preferences described is 'completely' unsafe to manipulate.*

From geometric consideration it is easy to understand that for three alternatives no strategic undershooting for scoring rules is possible. Example below presents an example of strategic undershooting for Borda with five alternatives.

Example 4 (Strategic undershooting under the Borda rule)**.** *Suppose 41 voters are using the Borda rule to select one of five alternatives. Let the number of different voter types present at the profile of sincere preferences be given by the following table:*

Preference order	*Number of voters*
$BCADE$	15
$CABED$	14
$CEDBA$	2
$DABEC$	10

When all voters vote honestly, A scores 102, B 110, C 109, D 59, and E 30; B wins. If between two and six DABEC types vote ADEBC, ceteris paribus, then C wins. If eight or more DABEC types vote ADEBC, ceteris paribus, then A wins. So DABEC voters can strategically undershoot at the profile of sincere preferences.

Strategic undershooting and overshooting are also possible under plurality. The examples that we have do, however, rely upon the tie-breaking procedure adopted.

4 Implications for the Gibbard and Satterthwaite theorem.

Proposition 1. *Suppose, at a profile R, a voter i can make a safe strategic vote L. Then there exists another profile Q, where the voter can make a safe strategic vote L and* $F(Q_{-i}(L) \succ_i F(Q)$.

Proof. Let i be a voter, and define V to be the set of all voters with preferences identical to those of i at R. Suppose that there exists a subset $U \subseteq V$ such that $i \in U$, and a linear order $L \neq R_i$ such that:

- every voter in U has an incentive to strategically vote L, and

- $F(R_{-U}(L)) \succ_V F(R)$,
- for every subset $W \subseteq V$ such that $W \supset U$ we have $F(R_{-W}(L)) \succeq_V F(R)$.

Without loss of generality we may assume that U is the smallest subset withe above properties relative to the set-theoretic inclusion. Then $F(R) = F(R_{U'}(L))$ for every subset $U' \subset U$. Let $U_1 = U - \{i\}$. Then voter i can make a safe strategic vote and is pivotal at $Q = R_{-U_1}(L)$. Indeed,

$$F(Q_{-i}(L) = F(R_{-U}(L)) \succ_V F(R) = F(Q).$$

since $U_1 \subset U$. □

From This and Theorem 1 we immediately deduce:

Theorem 2 (Extention of the GS Theorem)**.** *Suppose that the number of alternatives is at least three. Then any onto and non-dictatorial social choice rule is safely manipulable by a single voter.*

5 Conclusion and Further Research

This paper has formally distinguished between safe and unsafe manipulation of voting rules. Examples of unsafe manipulations were presented. The Gibbard-Satterthwaite theorem was extended to show that all onto, non-dictatorial social choice rules are safely manipulable.

The main two questions that stem from this research.

1. Under a reasonable probability distribution of votes (say, IC or IAC) what is the probability that someone can make a safe strategic vote?

Conitzer and Sandholm [4] observed that, for every election system for which the winner problem is in P, if the voters are unweighted and there are a fixed number m of candidates then the manipulation problem is in P. This result holds because a manipulator can easily evaluate all possible $m!$ manipulations. It is more difficult for her to decide whether or not she can make a safe strategic vote since she has to examine an exponential number (in the number of voters n) of subsets of like-minded voters who may join her.

2. Let the number of alternatives be fixed. For various social choice rules is the safe strategic vote problem in P?

We focused on social choice rules for two reasons. Firstly, it allowed us to formally introduce and illustrate the concepts of over and undershooting relatively simply. Secondly, our main result applies only to social choice rules. But the difference between safe and unsafe strategic votes is applicable to a much wider class of choice rules. For example, Slinko and White (2006) identified that strategic over and undershooting can occur under systems of proportional representation. Future research might consider the over and undershooting phenomena in other settings.

References

[1] Berg S and Lepelley D (1994) On Probability models in voting theory. Statistica Neerlandica **48**, 133146

[2] Bartholdi JJ, Tovey CA, and Trick MA (1989) The Computational Difficulty of Manipulating an Election: Social Choice and Welfare **6**:227–241

[3] Bartholdi JJ and Orlin JB (1991) Single Transferable Vote Resists Strategic Voting: Social Choice and Welfare **8**:341–354

[4] Conitzer V and Sandholm T (2002) Complexity of manipulating elections with few candidates. In: Proceedings of the 18th National Conference on Artificial Intelligence, pp. 314–319. AAAI Press.

[5] Conitzer V and Sandholm T (2003) Universal Voting Protocol Tweaks to Make Manipulation Hard. In Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03), pp. 781–788.

[6] Conitzer V and Sandholm T (2006) Nonexistence of Voting Rules That Are Usually Hard to Manipulate. In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06), 627– 634.

[7] Conitzer V, Sandholm T, and Lang J (2007) When Are Elections with Few Candidates Hard to Manipulate? Journal of the ACM **54**:1–33.

[8] Elkind E and Lipmaa H. (2005) Small coalitions cannot manipulate voting. In Proceedings of the 9th International Conference on Financial Cryptography and Data Security, pp. 285–297. Springer-Verlag Lecture Notes in Computer Science #3570.

[9] Faliszewski P., Hemaspaandra E, Hemaspaandra L, Rothe J (2006) A Richer Understanding of the Complexity of Election Systems. Arxiv preprint cs.GT/0609112 - arxiv.org

[10] Gibbard A (1973) Manipulation of Voting Schemes: A General Result: Econometrica **41**:587–601.

[11] Lepelley D and Mbih B (1994) The vulnerability of four social choice functions to manipulation of preferences: Social Choice and Welfare **11**:253-263

[12] Pritchard G and Slinko A (2006) On the Average Minimum Size of Manipulating Coalition, Social Choice and Welfare **27(2)**: 263–277.

[13] Pattanaik PK (1975) Strategic Voting Without Collusion under Binary and Democratic Group Decision Rules: The Review of Economic Studies **42**: 93–103.

[14] Satterthwaite MA (1975) Strategy-proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions: Journal of Economic Theory **10**:187–217.

[15] Slinko A (2002) The Asymptotic Strategy-proofness of the Plurality and the Run-off Rules, Social Choice and Welfare, **19**: 313–324.

[16] Slinko A (2002) On Asymptotic Strategy-Proofness of Classical Social Choice Rules, Theory and Decision **52**: 389–398.

[17] Slinko A (2004) How Large Should a Coalition Be to Manipulate an Election? Mathematical Social Sciences, **47(3)**: 289–293.

[18] Slinko A (2006) How the Size of a Coalition Affects its Chances to Influence an Election, Social Choice and Welfare, **26(1)**: 143–153.

[19] Slinko A and White S (2006) On the manipulability of proportional representation. Universite de Montreal. Departement de sciences economiques. Cahiers de recherche 2006-20.

[20] Slinko A and White S (2008) Is it ever safe to vote strategically? Report Series N 509. Department of Mathematics, The University of Auckland. 25pp.

Arkadii Slinko
Department of Mathematics
The University of Auckland
Private Bag 92019, Auckland, New Zealand
Email: a.slinko@auckland.ac.nz

Shaun White
The University of Auckland
student,
PO Box 5476, Wellesley St.,
Auckland 1141, NEW ZEALAND
Email: shaunwhite5476@hotmail.com

On the Agenda Control Problem for Knockout Tournaments

Thuc Vu, Alon Altman, Yoav Shoham

Abstract

Knockout tournaments are very common in practice for various settings such as sport events and sequential pairwise elimination elections. In this paper, we investigate the computational aspect of tournament agenda control, i.e., finding the agenda that maximizes the chances of a target player winning the tournament. We consider several modelings of the problem based on different constraints that can be placed on the structure of the tournament or the model of the players. In each setting, we analyze the complexity of finding the solution and show how it varies depending on the modelings of the problem. In general, constraints on the tournament structure make the problem become harder.

1 Introduction

Tournaments constitute a very common social institution. Their best known use is in sporting events, which attract millions of viewers and billions of dollars annually. But tournaments also play a key role in other social and commercial settings, ranging from the employment interview process to patent races and rent-seeking contests (see [12, 15, 9] for details).

Tournaments constitute a strict subclass of all competition formats, and yet they still allow for many different variations. All tournaments consist of *stages* during which several *matches* take place, matches whose outcome determines the set of matches in the next stage, and so on, until some final outcome of the tournament is reached. But tournaments vary in how many stages take place, which matches are played in each stage, and how the outcome is determined.

In this paper we focus on a narrower class of tournaments: *knockout* tournaments. In this very familiar format the players are placed at the leaf nodes of a binary tree. Players at sibling nodes compete against each other in a pairwise match, and the winner of the match moves up the tree. The player who reaches the root node is the winner of the tournament. We show an example in Figure 1.

The knockout tournament is not only a very popular tournament type used in practice for sporting events, but it is also a very common voting procedure. In the voting literature, it is referred to as sequential elimination voting with pairwise comparison [3, 8]. In this setting, each candidate is a player in the tournament, and the result of a match is based on the result of the pairwise comparison between the candidates as manifested in the electorate's votes. The result of the comparison can be either deterministic (as in [7]) or probabilistic (as in [11]).

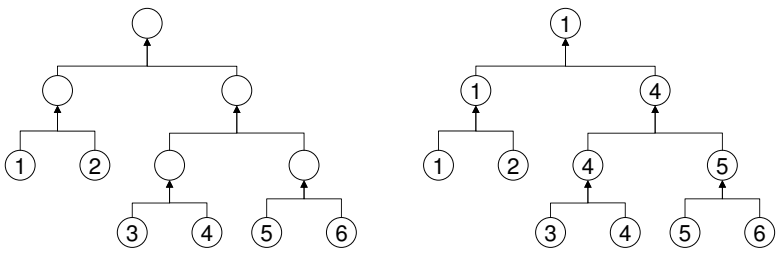


Figure 1: An example of a tournament agenda for 6 players and one possible outcome

In this paper we investigate the computational aspect of controlling the tournament agenda by the tournament designer. Specifically, we study the problem of how to find a tournament agenda to maximize the chance of a given target player winning the tournament. This seemingly simple question turns out to be surprisingly subtle and some of the answers are counter-intuitive. To begin with, note that the number of possible agenda grows extremely quickly with the number of players. Moreover, there are several variations of the problem modeling which lead to very different characteristics of the solutions. In particular, when there are no constraints on the modeling of the problem, it is unknown whether there exists an efficient algorithm to find the optimal agenda. However, when we place certain natural constraints on the structure of the tournament and the match results between the players, the problem becomes either easy or provably hard.

In Section 2 we summarize the existing results from the literature. After that we first discuss the general model for tournament design in Section 3. We then describe in Section 4 and 5 different constraints that can be placed on the model and our results for these settings. We conclude in Section 6 and suggest possible directions for future work.

2 Related Works

There are two approaches to determining the quality of a tournament design. The first one is axiomatic. For example, in [14], three axioms are proposed to specify what a "good" seeding should satisfy, called "Delayed Confrontation", "Sincerity Rewarded", and "Favoritism Minimized". Alternatively, in [6], a "Monotonicity" property is put forward.

In the second approach, rather than placing axiomatic constraints on the design, the goal is to optimize a certain quantity. The most common objective function is to find the design that maximizes the winning probability of the best player. This probability is also called the predictive power of the tournament. This has been the focus of much work (see, e.g., [5, 1, 13]). They all make the assumption that there is an ordering of the players based on their intrinsic abilities. In their models, the probability of one player winning against another is also known and is monotonic with regard to the abilities of the players, i.e., any player will have a higher chance of winning against a weaker player than winning against a stronger player. Nevertheless, they focus on only one type of tournament - balanced knockout tournaments, and only with small cases of n. In our work, we generalize the objective function to maximizing the winning probability of any given player, and we also consider various other modelings of the problem.

Tournament design problems are also addressed under the context of voting. In [7], the candidates are competing in an election based on sequential majority comparisons along a binary voting tree. In each comparison, the candidate with more votes wins and moves on; the candidate with less votes is eliminated. Essentially, the candidates are competing in a knockout tournament in which the result of each match is deterministic. The probability of winning a match is either 0 or 1. In this setting, without any constraints on the structure of the voting tree, there is a polynomial time algorithm to decide whether there exists a voting tree that will allow a particular candidate to win the election. The problem of finding the right voting tree (referred to as the "control" problem) is also addressed in [11] but with probabilistic comparison results instead. Here, the objective becomes finding a voting tree that allows a candidate to win the election with probability at least a certain value. In both [7, 11], the authors show that when the voting tree has to be balanced, some modified versions of the control problem are NP-complete.

The computational aspects of other methods of controlling an election are also considered in [2, 4]. Here, the organizer of the election is trying to change the result of the election through controls (such as adding or deleting) of the voters or candidates. It has been shown that for certain voting protocols, some methods of control are computationally hard to perform.

3 The general model and problem

We start out with the most general model of a knockout tournament. In this setting, there is no constraint on the structure of the tournament, as long as it only allows pairwise matches between players. We also assume that for any pairwise match, the probability of one player winning against the other is known. This probability can be obtained from past statistics or from some learning models. Here we do not place any constraints on the probabilities besides the fundamental properties. Thus there might be no transitivity between the winning probabilities, e.g., player i has more than 50% chance of beating player j, player j has more than 50% chance of beating player k, but player k also has more than 50% chance of beating player i.

We define a knockout tournament as the following:

Definition 1 (General Knockout Tournament)**.** *Given a set N of players and a matrix P such that $P_{i,j}$ denotes the probability that player i will win against player j in a pairwise elimination match and $0 \leq P_{i,j} = 1 - P_{j,i} \leq 1$ $(\forall i, j \in N)$, a knockout tournament $KT_N = (T, S)$ is defined by:*

- *A tournament structure T which is a binary tree with $|N|$ leaf nodes*
- *A seeding S which is a one-to-one mapping between the players in N and the leaf nodes of T*

We write KT_N as KT when the context is clear.

To carry out the tournament, each pair of players that are assigned to sibling leaf nodes with the same parent compete against each other in a pairwise elimination match. The winner of the match then "moves up" the tree and then competes against the winner of the other branch. The player who reaches the root of the tournament tree wins the tournament.

Intuitively the probability of a player winning the tournament depends on the probability that it will face a certain opponent and win against that opponent. We formally define this quantity below:

Definition 2 (Probability of Winning Tournament)**.** *Given a set N of players, a winning probability matrix P, and a knockout tournament $KT_N = (T, S)$, the probability of player k winning the tournament KT_N, denoted $q(k, KT_N)$ is defined by the following recursive formula:*

1. If $N = \{j\}$, then $q(k, KT_N) = \begin{cases} 1 & \text{if } k = j \\ 0 & \text{if } k \neq j \end{cases}$

2. If $|N| \geq 2$, let $KT_{N_1} = (T_1, S_1)$ and $KT_{N_2} = (T_2, S_2)$ be the two sub-tournaments of KT such that T_1 and T_2 are the two subtrees connected to the root node of T, and N_1 and N_2 are the set of players assigned to the leaf nodes of T_1 and T_2 by S_1 and S_2 respectively. If $k \in N_1$ then

$$q(k, KT_N) = \sum_{i \in N_2} q(k, KT_{N_1}) * q(i, KT_{N_2}) * P_{k,i}$$

and symmetrically for $k \in N_2$.

This recursive formula also gives us an efficient way to calculate $q(k, KT)$:

Proposition 1. *Given a set N of players, a winning probability matrix P, and a knockout tournament $KT_N = (T, S)$, the complexity of calculating $q(k, KT)$ with $k \in N$ is $O(|N|^2)$.*

Proof. First note that the number of operations is linear in the number of pairs (i, j) with $i, j \in N$ we consider. Moreover, for a given $i, j \in N$ we match up i and j only once. Thus the complexity is $O(|N|^2)$. □

Given a set of players N and the winning probabilities P between the players, the goal of the tournament designer is to come up with the tournament structure T and the seeding S that will maximize the probability of a given player $k \in N$ winning the tournament. This optimization

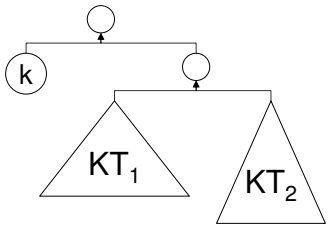


Figure 2: Biased knockout tournament KT' that maximizes the winning chance of k

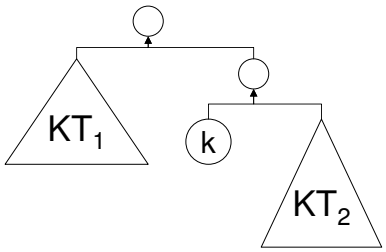


Figure 3: Knockout tournament KT

problem has a decision version which asks if there exist T and S such that the probability of k winning the tournament is greater than a given value θ.

The first intuition for the optimization problem is that the later any player plays in the tournament, the better chance she has of winning the tournament. We state and prove this intuition in the following proposition.

Proposition 2. *Given a set of players N and the winning probability matrix P, the tournament agenda that maximizes the probability of player $k \in N$ has the biased structure as in Figure 2 in which k has to play only the final match.*

Proof. We will prove this theorem by induction.

Base case: When $|N| = 2$, there is only one possible binary tree with 2 leaf nodes.

Inductive step: Assume that the theorem holds for N with $|N| \leq n-1$. For any given $k \in N$, we will show that it also holds for N with $|N| = n$ by converting any tournament agenda that does not have a biased structure to one with the same structure as in Figure 2 such that in this new tournament, k has at least the same chance of winning.

Let's consider any given tournament agenda KT that does not have the biased structure. Let KT_1 and KT_2' be the two disjoint sub-tournaments that make up KT, and let N_1, N_2' be the set of players assigned to KT_1, KT_2' respectively. Assume wlog that $k \in N_2'$. Since $|N_2'| < |N|$, the chance of k winning the tournament is maximized when KT_2' has the biased structure. Therefore we just need to compare the chance of k winning in KT (as shown in Figure 3) with its chance in KT' as shown in Figure 2:

$$q(k, KT) = \sum_{i \in N_2'} [p_{k,i} * q(i, KT_2)] * \sum_{j \in N_1} [p_{k,j} * q(j, KT_1)]$$

$$q(k, KT') = \sum_{j \in N_1, i \in N_2} p_{k,i} * q(i, KT_2) * p_{i,j} * q(j, KT_1) + \sum_{j \in N_1, i \in N_2} p_{k,j} * q(j, KT_1) * p_{j,i} * q(i, KT_2)$$

$$q(k, KT') - q(k, KT) = \sum_{j \in N_1, i \in N_2} q(j, KT_1) q(i, KT_2) \left[p_{k,j} * p_{j,i} + p_{k,i} * p_{i,j} - p_{k,i} * p_{k,j}\right]$$

$p_{i,j} + p_{j,i} = 1 \Rightarrow p_{k,j} * p_{j,i} + p_{k,i} * p_{i,j} \geq min\{p_{k,i}, p_{k,j}\} \geq p_{k,i} * p_{k,j}$
Therefore we have $q(k, KT') \geq q(k, KT)$.

□

Even though we know something about the shape of the optimal tournament structure, it is still an open problem whether there exists an efficient algorithm to find the exact optimal structure. However, when there are certain natural constraints on the structures of the tournament or the winning probabilities of the players, we manage to get a better analysis of the problem. In the next sections we will discuss the constraints and the results in the new settings.

4 A constraint on the structure of the tournament

In the previous section, we have shown that the optimal tournament structure is very unbalanced with the target player on one side and the rest of the players on the other side of the tree. One might say that this structure is unfair since the target player will have to compete only in the final match. One particular way to enforce fairness is to require the tournament structure to be a balanced binary tree (for simplicity, we assume that the number of players is a power of 2). This way, every player has to play the same number of matches in order to win the tournament.

Definition 3 (Balanced Knockout Tournament)**.** *Given a set N of players such that $|N| = 2^k$, a knockout tournament $KT = (T, S)$ is called balanced when T is a balanced binary tree.*

The balanced knockout tournament is in fact the most commonly used knockout tournament format in practice. In this setting, since the structure of the tournament is fixed, the remaining control of the tournament designer is in the seeding of the tournament, i.e., the assignment of players to the leaf nodes of the tree. Thus our previous problem is reduced to finding the seeding that will maximize the winning probability of a particular player. We have the following hardness result for the decision version of this problem:

Theorem 3. *Given a set N of players such that $|N| = 2^k$, and a winning probability matrix P, it is NP-complete to decide whether there exists a balanced knockout tournament KT such that $q(k, KT) \geq \delta$ for a given δ and $k \in N$.*

This theorem follows from Theorem 5. Therefore we will defer the discussion of the proof of this theorem to the next section. Note that the same result holds when the number of players is not a power of 2. In this case, when there is an odd number of players at any round, the tournament designer can let any player advance to the next round without competing. This allows certain bias, e.g., if the number of players is $2^k + 1$, then the target player can actually advance straight to the final match. Nevertheless, it is still NP-hard to find the optimal agenda for the target player.

5 Constraints on the player model

Besides placing a constraint on the structure of the tournament, we can also enforce certain constraints on the winning probabilities between the players. One such constraint can be on the possible values that the probabilities can take. Another constraint is a certain overall structure that the probabilities need to satisfy. We will discuss both types of constraints below.

5.1 Win-Lose Match Results

The first constraint we consider is requiring the result of each match to be deterministic, i.e., winning probabilities can only be either 0 or 1. As mentioned in Section 2, a knockout tournament in this setting is analogous to a sequential pairwise elimination election. Given a tournament agenda, a player in the tournament will either win the tournament for certain (winning with probability 1) or will lose for certain (winning with probability 0). Note that the winning probability matrix can be any arbitrary binary matrix.

When there is no constraint on the structure of the tournament, as shown in [7], there exists a polynomial time algorithm to find the tournament agenda that allows a given player k to win the tournament or decide that it is impossible for k to win. When the tournament has to be balanced, it is still an open problem.

We shall now discuss another problem model that we believe will be helpful for the understanding of the proof of Theorem 5. In this model, there is no constraint to the tournament tree, except that each player has to start from a pre-specified round. In other words, the tournament can take the shape of any binary tree, but each player has to start at certain depth of the tree.

Definition 4 (Knockout Tournament with Round Placements)**.** *Given a set N of players and a winning probability matrix P, a vector $R \in \mathbb{N}^{|N|}$, if there exists a knockout tournament KT such that in KT, player i starts from round R_i (the leaf nodes with the maximum depth in the tree are considered to be at round 0), then R is called a* feasible *round placement and such tournament KT is called a knockout tournament with round placement R. When there is an odd number of players at any given round, one player playing at that round can automatically advance to the next round.*

Note that when all players have round placement 0, the tournament is balanced. We have the following hardness result:

Theorem 4. *Given a set of players N, the winning probability matrix P such that $\forall i \neq j \in N$, $P_{i,j} \in \{0, 1\}$, and a feasible round placement R, it is NP-complete to decide whether there exists a tournament agenda KT with round placements R such that $q(k, KT) \geq \delta$ for a given δ and $k \in N$.*

Proof. It is easy to show that the problem is in NP. We will show the problem is NP-complete using a reduction from the Vertex Cover problem.

Vertex Cover: Given a graph $G = \{V, E\}$ and an integer k, is there a subset $C \in V$ such that $|C| \leq k$ and C covers E?

Reduction method:
We construct a tournament $KT = (T, S)$ with a special player o such that o wins KT with probability 1 if and only if there exists a vertex cover of size at most k.

KT contains the following players[1]:

1. Objective player: o which starts at round 0.
2. Vertex players: $\{v_i \in V\}$ which start at round 0. There are $n = |V|$ such players.
3. Edge players: $\{e_i \in E\}$. There are $m = |E|$ such players. e_i starts at round $(n - k + i - 1)$.
4. Filler players: For each round r such that $(n-k+m) > r \geq (n-k)$, there is one filler player f^r that starts at round r. Thus there are a total of m of them. They are meant for player o.
5. Holder players: For each round r, there are a set of holder players h_i^r (i.e., multiple copies of h^r) that start at round r. These players are meant for the vertex players. The number of copies of h^r depends on the value of r:
   - If $0 \leq r < (n - k)$, there are $(n - r - 1)$ of them
   - If $(n - k) \leq r < (n - k + m)$, there are $(k - 1)$ of them
   - If $(n - k + m) \leq r < (n + m)$, there are $(m + n - r - 1)$ of them

The winning probabilities between the players are assigned as in Table 5.1. In a nutshell:

1. o only wins against v_i and f with probability 1 (always wins) and loses against all others with probability 1 (always loses).
2. v_i always wins against h^r, e_j that it covers, and $v_{i'}$ with $i' > i$. It always loses against all other players.
3. e_j always wins against h^r, f, $e_{j'}$ with $j' > j$.
4. Between two f^r players, the winner can be either one of them.
5. Between two h^r players, the winner can be either one of them.

	o	v_j	e_j	f^r	h_i^r
o	-	1	0	1	0
v_i	-	1 if $i \leq j$, 0 otherwise	1 if v_i covers e_j, 0 otherwise	0	1
e_i	-	-	1 if $i \leq j$, 0 otherwise	1	1
f^r	-	-	-	arbitrary	1
h_i^r	-	-	-	-	arbitrary

Table 1: The winning probabilities of row players against column players in KT

The reduction is polynomial since the numbers of players in the tournament is polynomial.

We first need to show how to construct an agenda KT that lets o win with probability 1 if there exists a vertex cover C of size at most k. The desired KT is composed of three phases:

Phase 1: Phase 1 is the first $(n-k)$ rounds. In this phase, we eliminate all vertex players that are not in C while keeping the remaining vertex players, and o. At each round r, match up o with $v' \notin C$ and let each of the $(n-r)$ holder players h^r match up with the remaining v_i. Notice that after each round, one vertex player gets eliminated and there is one less h^r. After $(n-k)$ rounds, there are k vertex players left. Each of them corresponds to a vertex in C.

Phase 2: Phase 2 is the following m rounds. In this phase, we eliminate all edge players. For each round, we match up o with f^r. At each round r, there will be one edge player e starting at that round. We match e against $v_i \in C$ that covers it. For the remaining vertex players, we match them up with $(k-1)$ holder players h^r. After m rounds, all of the edge players will be eliminated (since the k vertex players left form a vertex cover). The remaining players at the end of this phase are k vertex players and o.

Phase 3: Phase 3 is the final k rounds after Phase 2. In this phase, we eliminate the remaining vertex players. At each round, the number of new holder players starting at that round is one less than the number of remaining vertex players. We match up the vertex players with h^r, and o with the remaining v. At the end of this phase, only o remains.

For the other direction, we need to prove that o can win the tournament with probability 1 only if there is a vertex cover C of size k. First note that during Phase 1, for o not to get eliminated, it has to play against a vertex player v. Thus after the first $(n-k)$ rounds, there are at most k vertex players remaining (there can be less if two vertex players play against each other).

During Phase 2, the only ways that an edge player e can be eliminated is to play against v that covers it or play against another edge player e' which started at an earlier round. If e is eliminated by e', there must be either one h^r or one v that was eliminated earlier by an edge player e'' (which can possibly be e'). Since there is only $(k-1)$ holder players at each round, if h^r was eliminated by e'', two vertex players must have played against each other and one of them must have been eliminated. Thus for both cases, there is at least one v that got eliminated. Note that in this phase, at any round, there are only $(k+1)$ new players. Therefore, at the end of this phase, there are exactly $(k+1)$ players remaining including o. If all edge players get eliminated by vertex players, there are k vertex players remaining. If there is at least one e which did not get eliminated or got eliminated by another e', there are less than k vertex players remaining.

Now during Phase 3, for o to win the tournament, o can only play against a vertex player. Thus the number of vertex players is reduced each round by 1. Moreover, since there are $(k-1)$ holder players h^r starting at the first round of the phase, and one less for each round after that, if there are less than k vertex players at the beginning of Phase 3, there will be at least one non-vertex player remaining. If that is the case, at the last round of Phase 3, there must be at least one edge or holder player remaining and o will lose the tournament.

Therefore, for o to win the tournament, there must be k vertex players at the beginning of Phase 3. This implies all edge players must have been eliminated by vertex players during Phase 2. So

[1]We overload some notations here but the given the context, it should be clear

each edge player must be covered by at least one of the remaining vertex players after Phase 1. Since there are at most k of them after Phase 1, these remaining vertex players form a vertex cover of size at most k. □

After placing this constraint on the structure of the tournament tree, the tournament design problem has changed from easy to hard. This gives an indication that the design problem for balanced knockout tournament with deterministic match results is probably also hard.

5.2 Win-Lose-Tie Match Results

When the match results are deterministic, it is an open problem whether there exists an efficient algorithm to find the optimal balanced knockout tournament for a given player. Surprisingly, when we allow there to be a tie between two players (each has equal chance of winning), the problem becomes provably hard.

Theorem 5. *Given a set of players N, a winning probability matrix P such that $P_{i,j} \in \{0, 1, 0.5\}$, it is NP-complete to decide whether there exists a balanced knockout tournament KT such that $q(k, KT) \geq \delta$ for a given δ and $k \in N$.*

The proof of Theorem 5 is similar to the proof of Theorem 4 with two modifications to the reduction:
1. We need to construct some gadgets that simulate the round placements, i.e., if player i starts from round r, player i will not be eliminated until round r. In order to achieve this, we will introduce $(2^r - 1)$ filler players that only player i can beat. This will keep player i busy until at least round r
2. We need to make sure that the round placement for any player is at most $O(log(n))$ with n equal to the size of the Vertex Cover Problem so that the size of the tournament is still polynomial.

Proof of Theorem 5. Similar to the Proof of Theorem 4, we show here a reduction from the Vertex Cover problem.

Reduction method:
We construct a tournament $KT = (T, S)$ with a special player o such that o wins KT with probability 1 if and only if there exists a vertex cover of size at most k.

KT contains the following players:

1. Objective player: o
2. Vertex players: $\{v_i \in V\}$ and an extra special vertex v_0 which does not cover any edge. If we let $n = |V|$ then there are $n + 1$ vertex players.
3. Edge players: $\{e_i \in E\}$. There are $m = |E|$ edge players.
4. Filler players: For each round r such that $0 < r \leq \lceil log(n - k) \rceil$, there are k filler players $f^r_{v,i}$, i.e., there are k copies of f^r_v. These players are meant to keep at least k vertex players advancing to the next round. For each round r such that $\lceil log(n - k) \rceil < r \leq \lceil log(n - k) \rceil + \lceil log(m) \rceil$, there are k filler players $f^r_{e,i}$. These are meant for the edge players. We might refer to both types of filler players as f^r_i or just simply f^r.
5. Holder players: For each edge player e_i, there are $2^{\lceil log(n-k) \rceil} - 1$ edge holder players $h^l_{e_i}$. These will make sure no edge player will be eliminated before reaching round $\lceil log(n - k) \rceil + 1$. For each filler player f^r_i, there are $2^r - 1$ holder players $h^l_{f^r_i}$ that will make sure no filler player will be eliminated before reaching round r. There are also

$K = 2^{\lceil log(n-k) \rceil + \lceil log(m) \rceil + \lceil log(k+1) \rceil + 1} - 1$ special holder players h_o^l that will allow player o to advance to the final match.

The winning probabilities between the players are assigned as in Table 5.2. In a nutshell:

1. o only wins against v_i and h_o with probability 1 (always wins) and loses against all others with probability 1 (always loses).
2. v_i always wins against f^r (both f_v^r and f_e^r), e_j that it covers, and $v_{i'}$ with $i' > i$. It always loses against all other players. The special vertex player v_0 does not win against any edge player but wins against any other vertex player.
3. e_j always wins against h_{e_j}, f^r, and wins with probability 0.5 against another $e_{j'}$.
4. For the holder players, each of them only loses to the player it is meant for. For example, edge holder player $h_{e_j}^l$ only loses to the edge player e_j. Holder players tie when playing against each other or against an edge player (except for the edge holder players they are meant for). They always win against o and vertex players.

	v_j	e_j	$f_j^{r'}$	$h_{e_j}^{l'}$	$h_{f_j^{r'}}^{l'}$	$h_o^{l'}$
o	1	0	0	0	0	1
v_i	1 if $i < j$, 0 otherwise	1 if v_i covers e_j, 0 otherwise	1	0	0	0
e_i	-	0.5	0.5	1 if $i = j$, 0.5 otherwise	1	1
f_i^r	-	-	0.5	0.5	1 if $f_i^r = f_j^{r'}$, 0.5 otherwise	1
$h_{e_i}^l$	-	-	-	0.5	0.5	1
$h_{f_i^r}^l$	-	-	-	-	0.5	1
h_o^l	-	-	-	-	-	0.5

Table 2: The winning probabilities of row players against column players in KT

The reduction is polynomial since the number of players in the tournament is $O(K)$. Without loss of generality, we assume that the number of total players is a power of 2 because we can always add more h_o players and this will not affect the reduction shown below. Note that we consider the first round as round 1.

First we need to show how to construct an agenda KT that let o win with probability 1 if there exists a vertex cover C of size at most k. The desired KT is composed of two phases:

Phase 1: Phase 1 is the first $\lceil log(n-k) \rceil$ rounds. In this phase, we eliminate all $v \notin C$ except the special vertex player v_0 while keeping o and all edge players. During this phase, for each player that has corresponding holder players, we will match them together. This will help each edge player e to get to round $\lceil log(n-k) \rceil + 1$, and each filler player f^r to get to round r. We also match o with the holder player h_o to help o advancing to the final round. At round $1 \leq r \leq \lceil log(n-k) \rceil$, if the vertex v_i is in C, we match the vertex player v_i with the filler player f^r. Otherwise we match it with another vertex player that is not in C. At the end of this phase , there are only $k+1$ vertex players remaining. One of them is the special vertex player v_0.

Phase 2: Phase 2 is the following $\lceil log(m) \rceil + \lceil log(k+1) \rceil + 1$ rounds. In this phase, we eliminate all the edge players by repeatedly matching each vertex player with the edge players that it covers. If there are more edge players than vertex players, we will match the remaining edge players who are covered by the same vertex player with each other. If there is any edge player that does not have a match, we will match it with the edge filler player f_e^r. Note that there are at most k edge players

that do not have a match. After each round, at least half of the edge players will be eliminated. Thus after at most $\lceil log(m) \rceil$ rounds, all the edge players will be eliminated. There will be only vertex players v, o and h_o remaining. We just need to match them up until o is the only player left since o wins against v and h_o with probability 1.

For the other direction, we need to prove that o can win the tournament with probability 1 only if there is a vertex cover C of size k. We need to show that if o wins with probability 1, after Phase 1, there will be only $k+1$ vertex players remaining including the special vertex v_0, and during Phase 2, all edge players will be eliminated by one of those remaining vertex players, i.e., no edge players gets eliminated during phase 1.

First note that for o to win with probability 1, no holder players except h_o can get to the final. Thus during the first $\lceil log(n-k) \rceil$ rounds, no edge player can get eliminated since there are $(2^{\lceil log(n-k) \rceil} - 1)$ holder players for each edge player. Also no filler player f^r can get eliminated before round r. At round r such that $r \leq \lceil log(n-k) \rceil$, the only way for a vertex player to advance to the next round is either playing against a filler player f^r or another vertex player. It cannot advance by playing against an edge player that it covers, since that would eliminate that edge player too early. It cannot advance by playing a filler player $f^{r'}$ with $r' > r$ either, since that would eliminate $f^{r'}$ too early. Therefore, besides k vertex players playing against f^r, at least half of the remaining vertex players will be eliminated after each round. At the end of round $\lceil log(n-k) \rceil$, there can be only at most $k+1$ vertex players remaining. Note that since vertex player v_0 wins against any other vertex players, it must still remain.

For o to win the tournament with probability of 1, o must not play against any edge players either. Moreover, note that when two edge players play against each other, each of them has a 50% chance moving on to the next round. Thus the only way that an edge player gets eliminated is to play against a vertex player that covers it. So, each edge player must be covered by at least one of the remaining k vertex players (since v_0 does not cover any edge). Thus the set of k remaining vertex players forms a vertex cover of size k. □

The problem of finding the optimal balanced knockout tournament with Win-Lose-Tie match results is reducible to our general balanced knockout tournament. This constitutes the proof for Theorem 3.

When there is no constraint on the structure of the tournament, there exists a polynomial time algorithm to find an agenda that will allow a target player to win with probability 1 or decide that such an agenda does not exist. This algorithm is a modification the algorithm introduced in [7] to compute possible winners. In this algorithm, when there is a tie between 2 players, we remove all the edges between them in the tournament graph. We will then proceed to finding all the winning paths from the target player to other players in the tournament. If there is a winning path to each of the other players, there exists an agenda to make the target player win, and that agenda is the binary tree formed by combining the winning paths.

5.3 Monotonic Winning Probabilities

Another natural constraint is to require a certain overall structure of the winning probability matrix P. One of the most common models in the literature is the monotonic model (see for example [5, 10, 6, 14]). In this model, the players are numbered from 1 to n in descending order of unknown intrinsic abilities. We only know the probability of one player winning against another. These winning probabilities are also correlated to the intrinsic abilities.

Definition 5 (Knockout Tournament with Monotonic Winning Probabilities)**.** *A knockout tournament $KT = \{N, T, S, P\}$ has monotonic winning probabilities when P satisfies the following constraints:*

1. $p_{ij} + p_{ji} = 1$

2. $p_{ij} \geq p_{ji} \qquad \forall (i,j) : i \leq j$

3. $p_{ij} \geq p_{i(j+1)} \qquad \forall (i,j)$

As in the case of deterministic match results, when we require the tournament to be balanced, the complexity of finding the optimal tournament is unknown. Unlike the case of balanced tournaments with win-lose-tie, the monotonicity condition is too restrictive. Yet, when we relax this condition to allow small violations, we can obtain a hardness result. We call the new condition ϵ-monotonicity.

Definition 6 (ϵ-monotonicity)**.** *A winning probability matrix P is ϵ-monotonic with $\epsilon > 0$ when P satisfies the following constraints:*

1. $p_{ij} + p_{ji} = 1$

2. $p_{ij} \geq p_{ji} \qquad \forall (i,j) : i \leq j$

3. $p_{ij} \geq p_{i(j')} - \epsilon \qquad \forall (i,j,j') : j' > j$

As ϵ goes to 0, the winning probability matrix P will gets closer to being monotonic. Note that we only relax the second requirement of monotonicity. In this setting, the problem of finding the optimal balanced agenda is provably hard:

Theorem 6. *Given a set of players N, a ϵ-monotonic winning probability matrix P with $\epsilon > 0$, it is NP-complete to decide if there exists a balanced knockout tournament KT such that $q(k, KT) \geq \delta$ for a given δ and $k \in N$.*

Proof. To prove this theorem, we show a reduction from the Vertex Cover Problem to the tournament design problem in this setting. The reduction is similar to the proof of Theorem 3 with the same set of players but with slightly different winning probabilities (shown in Table 5.3). Essentially, we convert the probabilities of a vertex player winning against edge players that it does not cover from 0 to $(1-\epsilon)$. Similarly for o, it now either wins with probabilities 1 as before or with probabilities $(1-\epsilon)$. The new winning probabilities are ϵ-monotonic with this ordering of players in descending strengths: o, v_0... v_n, e_1... e_m, f^r, h_e, h_{f^r}, h_o. Note that for o to win the tournament with probability 1, she can only play against v and h_o. Thus all players of other types must be eliminated with probability 1. This allows the proof of Theorem 3 to hold in this setting.

	v_j	e_j	$f_j^{r'}$	$h_{e_j}^{l'}$	$h_{f_j^{r'}}^{l'}$	$h_o^{l'}$
o	1	$1-\epsilon$	$1-\epsilon$	$1-\epsilon$	$1-\epsilon$	1
v_i	1 if $i<j$, 0 ow.	1 if v_i covers e_j, $(1-\epsilon)$ ow.	1	$1-\epsilon$	$1-\epsilon$	$1-\epsilon$
e_i	-	0.5	1	1 if $i=j$, $(1-\epsilon)$ ow.	1	1
f_i^r	-	-	0.5	0.5	1 if $f_i^r = f_j^{r'}$, 0.5 otherwise	1
$h_{e_i}^l$	-	-	-	0.5	0.5	1
$h_{f_i^r}^l$	-	-	-	-	0.5	1
h_o^l	-	-	-	-	-	0.5

Table 3: The ϵ-monotonic winning probabilities of row players against column players in KT

□

6 Conclusion and future work

In this paper we have investigated the computational aspect of agenda control for knockout tournaments. We have considered several modelings of the problem based on different constraints that can be placed on the structure of the tournament or the model of the players. In particular, we have shown that when the tournament has to be balanced, the agenda control problem is NP-hard, even when the match results can only be win, lose, or tie, or when the winning probabilities between the players have to be ϵ-monotonic. This suggests that it is hard to design a knockout tournament with maximum predictive power. When the match results are deterministic, the complexity of the control problem remains an open problem for future work. Other directions include finding optimal agenda for other objective functions such as fairness or "interestingness" of the tournament, or considering other constraints on the tournament structure and player models.

References

[1] D. R. Appleton. May the best man win? *The Statistician*, 44(4):529–538, 1995.

[2] J. Bartholdi, C. Tovey, and M. Trick. How hard is it to control en election? *Mathematical and Computer Modeling*, 16(8/9):27–40, 1992.

[3] S. J. Brams and P. C. Fishburn. Voting procedures. In K. J. Arrow, A. K. Sen, and K. Suzumura, editors, *Handbook of Social Choice and Welfare*.

[4] E. Hemaspaandra, L. A. Hemaspaandra, and J. Rothe. Anyone but him: The complexity of precluding an alternative. *Artif. Intell.*, 171(5-6):255–285, 2007.

[5] J. Horen and R. Riezman. Comparing draws for single elimination tournaments. *Operations Research*, 33(2):249–262, mar 1985.

[6] F. K. Hwang. New concepts in seeding knockout tournaments. *The American Mathematical Monthly*, 89(4):235–239, apr 1982.

[7] J. Lang, M. S. Pini, F. Rossi, K. B. Venable, and T. Walsh. Winner determination in sequential majority voting. In *IJCAI*, pages 1372–1377, 2007.

[8] J.-F. Laslier. *Tournament solutions and majority voting*. Springer, 1997.

[9] G. C. Loury. Market structure and innovation. *The Quarterly Journal of Economics*, 93(3):395–410, August 1979.

[10] J. W. Moon and N. J. Pullman. On generalized tournament matrices. *SIAM Review*, 12(3):384–399, jul 1970.

[11] S. K. N. Hazon, P. E. Dunne and M. Wooldridge. How to rig an election. In *The 9th Bar-Ilan Symposium on the Foundations of Artificial Intelligence*, 2007.

[12] S. Rosen. Prizes and incentives in elimination tournaments. *The American Economic Review*, 76(4):701–715, sep 1986.

[13] D. Ryvkin. The predictive power of noisy elimination tournaments. Technical report, The Center for Economic Research and Graduate Education - Economic Institute, Prague, Mar. 2005.

[14] A. J. Schwenk. What is the correct way to seed a knockout tournament? *The American Mathematical Monthly*, 107(2):140–150, feb 2000.

[15] G. Tullock. *Toward a Theory of the Rent-seeking Society*. Texas A&M University Press, 1980.

Thuc Vu, Alon Altman, Yoav Shoham
Department of Computer Science
Stanford University
Stanford, USA
Email: {thucvu,epsalon,shoham}@stanford.edu

Complexity of unweighted coalitional manipulation under some common voting rules

Lirong Xia, Vincent Conitzer, Ariel D. Procaccia, and Jeffrey S. Rosenschein

Abstract

In this paper, we study the computational complexity of the unweighted coalitional manipulation (UCM) problem under some common voting rules. We show that the UCM problem under maximin is NP-complete. We also show that the UCM problem under ranked pairs is NP-complete, even if there is only one manipulator. Finally, we present a polynomial-time algorithm for the UCM problem under Bucklin.

1 Introduction

Voting is a methodology for a group of agents (or voters) to make a joint choice from a set of alternatives. Each agent reports his or her preferences over the alternatives; then, a *voting rule* is applied to aggregate the preferences of the agents—that is, to select a winning alternative. However, sometimes a subset of the agents can report their preferences insincerely to make the outcome more favorable to them. This phenomenon is known as *manipulation*. A rule for which no group of agents can ever beneficially manipulate is said to be *group strategy-proof*; if no single agent can ever beneficially manipulate, the rule is said to be *strategy-proof* (a weaker requirement).

Unfortunately, any strategy-proof voting rule will fail to satisfy some natural property. The celebrated Gibbard-Satterthwaite theorem [10, 16] states that when there are three or more alternatives, there is no strategy-proof voting rule that satisfies non-imposition (for every alternative, there exist votes that would make that alternative win) and non-dictatorship (the rule does not simply always choose the most-preferred alternative of a single fixed voter). However, the mere existence of beneficial manipulations does not imply that voters will use them: in order to do so, voters must also be able to *discover* the manipulation, and this may be computationally hard. Recently, the approach of using computational complexity to prevent manipulation has attracted more and more attention. In early work [2, 1], it was shown that when the number of alternatives is not bounded, the second-order Copeland and STV rules are hard to manipulate, even by a single voter. More recent research has studied how to modify other existing rules to make them hard to manipulate [3, 7].

Some attention has been given to a problem known as *weighted coalitional manipulation* (WCM) in elections. In this setting, there is a coalition of manipulative voters trying to coordinate their actions in a way that makes a specific alternative win the election. In addition, the voters are weighted; a voter with weight k counts as k voters voting identically. Previous work has established that this problem is computationally hard under a variety of prominent voting rules, even when the number of candidates is constant [6, 11].

However, and quite surprisingly, the current literature contains few results regarding the *unweighted* version of the coalitional manipulation problem (UCM), which is in fact more natural in most settings. Recently, it has been shown that UCM is NP-complete under a family of voting rules derived from the Copeland rule, even with only two manipulators [8]. Zuckerman et al. [20] have established, as corollaries of their main theorems, that unweighted coalitional manipulation is tractable under the Veto and Plurality with Runoff voting rules.

In this paper, we study the computational complexity of the unweighted coalitional manipulation problem under the maximin, ranked pairs, and Bucklin rules. After briefly recalling basic notations and definitions, we show that the UCM problem under maximin is NP-complete for any fixed number of manipulators (at least two). We then show that the UCM problem under ranked

pairs is NP-complete, even when there is only one manipulator (just as this is hard for second-order Copeland and STV). Finally, we present a polynomial-time algorithm for the UCM problem under Bucklin.

2 Preliminaries

Let $\mathcal{C}$ be the set of *alternatives* (or *candidates*). A linear order on $\mathcal{C}$ is a transitive, antisymmetric, and total relation on $\mathcal{C}$. The set of all linear orders on $\mathcal{C}$ is denoted by $L(\mathcal{C})$. An n-voter profile P on $\mathcal{C}$ consists of n linear orders on $\mathcal{C}$. That is, $P = (R_1, \ldots, R_n)$, where for every $i \leq n$, $R_i \in L(\mathcal{C})$. The set of all profiles on $\mathcal{C}$ is denoted by $P(\mathcal{C})$. In the remainder of the paper, we let m denote the number of alternatives (that is, $|\mathcal{C}|$).

A *voting rule* r is a function from the set of all profiles on $\mathcal{C}$ to $\mathcal{C}$, that is, $r : P(\mathcal{C}) \to \mathcal{C}$. The following are some common voting rules studied in this paper.

1. *(Positional) scoring rules*: Given a *scoring vector* $\vec{v} = (v(1), \ldots, v(m))$, for any vote $V \in L(\mathcal{C})$ and any $c \in \mathcal{C}$, let $s(V, c) = v(j)$, where j is the rank of c in V. For any profile $P = (V_1, \ldots, V_n)$, let $s(P, c) = \sum_{i=1}^{n} s(V_i, c)$. The rule will select $c \in \mathcal{C}$ so that $s(P, c)$ is maximized. Two examples of scoring rules are *Borda*, for which the scoring vector is $(m-1, m-2, \ldots, 0)$, and *plurality*, for which the scoring vector is $(1, 0, \ldots, 0)$.

2. *Maximin*: Let $N_P(c_i, c_j)$ denote the number of votes that rank c_i ahead of c_j. The winner is the alternative c that maximizes $\min\{N_P(c, c') : c' \in \mathcal{C}, c' \neq c\}$.

3. *Bucklin*: An alternative c's Bucklin score is the smallest number k such that more than half of the votes rank c among the top k alternatives. The winner is the alternative that has the smallest Bucklin score. (Sometimes, ties are broken by the number of votes that rank an alternative among the top k, but for simplicity we will not consider this tie-breaking rule here.)

4. *Ranked pairs* [17]: This rule first creates an entire ranking of all the alternatives. $N_P(c_i, c_j)$ is defined as for the maximin rule. In each step, we consider a pair of alternatives c_i, c_j that we have not previously considered (as a pair): specifically, we choose the remaining pair with the highest $N_P(c_i, c_j)$. We then fix the order $c_i > c_j$, unless this contradicts previous orders that we fixed (that is, it violates transitivity). We continue until we have considered all pairs of alternatives (hence, in the end, we have a full ranking). The alternative at the top of the ranking wins.

All of these rules allow for the possibility that multiple alternatives end up tied for the win. Technically, therefore, they are really *voting correspondences*; a correspondence can select more than one winner. In the remainder of this paper, we will sometimes somewhat inaccurately refer to the above correspondences as rules. We will consider two variants of the manipulation problem: one in which the goal is to make the preferred alternative the unique winner, and one in which the goal is to make sure that the preferred alternative is among the winners. We study the *constructive* manipulation problem, in which the goal is to make a given alternative win.

Definition 1 *An* unweighted coalitional manipulation (UCM) *instance is a tuple* (r, P^{NM}, c, M)*, where* r *is a voting rule,* P^{NM} *is the non-manipulators' profile,* c *is the alternative preferred by the manipulators, and* M *is the set of manipulators.*

Definition 2 *The* UCM unique winner (UCMU) *problem is: Given a UCM instance* (r, P^{NM}, c, M)*, we are asked whether there exists a profile* P^M *for the manipulators such that* $r(P^{NM} \cup P^M) = \{c\}$*.*

Definition 3 *The* UCM co-winner (UCMC) *problem is: Given a UCM instance* (r, P^{NM}, c, M)*, we are asked whether there exists a profile* P^M *for the manipulators such that* $c \in rP^{NM} \cup P^M$*.*

3 Maximin

In this section, we show that the UCMU and UCMC problems under maximin are NP-complete, by giving a reduction from the *two vertex disjoint paths in directed graph* problem, which is known to be NP-complete [12].

Definition 4 *The* two vertex disjoint paths in directed graph *problem is: We are given a directed graph G and two disjoint pairs of vertices (u, u') and (v, v'), where u, u', v, v' are all different from each other. We are asked whether there exist two directed paths $u \to u_1 \to \ldots \to u_{k_1} \to u'$ and $v \to v_1 \to \ldots \to v_{k_2} \to v'$ such that $u, u', u_1, \ldots, u_{k_1}, v, v', v_1, \ldots, v_{k_2}$ are all different from each other.*

For any profile P and any pair of alternatives c_1, c_2, let $D_P(c_1, c_2)$ denote the number of times that c_1 is ranked higher than c_2 in P minus the number of times that c_2 is ranked higher than c_1 in P. That is,

$$D_P(c_1, c_2) = |\{R \in P : c_1 \succ_R c_2\}| - |\{R \in P : c_2 \succ_R c_1\}|$$

The next lemma has previously been used by others [13, 4].

Lemma 1 *Given a profile P and $F : \mathcal{C} \times \mathcal{C} \to \mathbb{Z}$ such that*

1. *for all $c_1, c_2 \in \mathcal{C}$, $c_1 \neq c_2$, $F(c_1, c_2) = -F(c_2, c_1)$, and*
2. *either for all pairs of alternatives $c_1, c_2 \in \mathcal{C}$ (with $c_1 \neq c_2$), $F(c_1, c_2)$ is even, or for all pairs of alternatives $c_1, c_2 \in \mathcal{C}$ (with $c_1 \neq c_2$), $F(c_1, c_2)$ is odd,*

there exists a profile P such that for all $c_1, c_2 \in \mathcal{C}$, $c_1 \neq c_2$, $D_P(c_1, c_2) = F(c_1, c_2)$ and $|P| \leq \frac{1}{2} \sum_{c_1, c_2 : c_1 \neq c_2} |F(c_1, c_2) - F(c_2, c_1)|$.

Theorem 1 *The UCMU and UCMC problems under maximin are NP-complete for any fixed number of manipulators (as long as it is at least 2).*

Proof of Theorem 1: It is easy to verify that the UCMU and UCMC problems under maximin are in NP. We first show that UCMU is NP-hard, by giving a reduction from the two vertex disjoint paths in directed graph problem. Let the instance of the two vertex disjoint paths in directed graph problem be denoted by $G = (V, E)$, (u, u') and (v, v') where $V = \{u, u', v, v', c_1, \ldots, c_{m-5}\}$. Without loss of generality, we assume that every vertex is reachable from u or v (otherwise, we can remove the vertex from the instance). We also assume that $(u, v') \notin E$ and $(v, u') \notin E$ (since such edges cannot be used in a solution). Let $G' = (V, E \cup \{(v', u), (u', v)\})$, that is, G' is the graph obtained from G by adding (v', u) and (u', v). We construct a UCMU instance as follows.
Set of alternatives: $\mathcal{C} = \{c, u, u', v, v', c_1, \ldots, c_{m-5}\}$.
Alternative preferred by the manipulators: c.
Number of unweighted manipulators: $|M|$ (for some $|M| \geq 2$).
Non-manipulators' profile: P^{NM} satisfying the following conditions:

1. For any $c' \neq c$, $D_{P^{NM}}(c, c') = -4|M|$.
2. $D_{P^{NM}}(u, v') = D_{P^{NM}}(v, u') = -4|M|$.
3. For any $(s, t) \in E$ such that $D_{P^{NM}}(t, s)$ is not defined above, we let $D_{P^{NM}}(t, s) = -2|M| - 2$.
4. For any $s, t \in \mathcal{C}$ such that $D_{P^{NM}}(t, s)$ is not defined above, we let $|D_{P^{NM}}(t, s)| = 0$.

The existence of such a P^{NM}, whose size is polynomial in m, is guaranteed by Lemma 1.

We can assume without loss of generality that each manipulator ranks c first. Therefore, for any $c' \neq c$, $D_{P^{NM} \cup P^M}(c, c') = -3|M|$.

We are now ready to show that $Maximin(P^{NM} \cup P^M) = \{c\}$ if and only if there exist two vertex disjoint paths from u to u' and from v to v' in G. First, we prove that if there exist such paths in G, then there exists a profile P^M for the manipulators such that $Maximin(P^{NM} \cup P^M) = \{c\}$. Let $u \to u_1 \to \ldots \to u_{k_1} \to u', v \to v_1 \to \ldots \to v_{k_2} \to v'$ be two vertex disjoint paths. Let $V' = \{u, u', v, v', u_1, \ldots, u_{k_1}, v_1, \ldots, v_{k_2}\}$. Then, because any vertex is reachable from u or v in G, there exists a connected subgraph G^* of G' (which still includes all the vertices) in which $u \to u_1 \to \ldots \to u_{k_1} \to u' \to v \to v_1 \to \ldots \to v_{k_2} \to v' \to u$ is the only cycle. Therefore, there exists a linear order O over $V \setminus V'$ such that for any $t \in V \setminus V'$, either 1. there exists $s \in V \setminus V'$ such that $s \succ_O t$ and $(s, t) \in E$, or 2. there exists $s \in V'$ such that $(s, t) \in E$. We let

$$\begin{aligned} P^M = &\{(|M| - 1)(c \succ u \succ u_1 \succ \ldots \succ u_{k_1} \succ u' \succ v \succ v_1 \succ \ldots \succ v_{k_2} \succ v' \succ O)\} \\ &\cup \{c \succ v \succ v_1 \succ \ldots \succ v_{k_2} \succ v' \succ u \succ u_1 \succ \ldots \succ u_{k_1} \succ u' \succ O\} \end{aligned}$$

Then, we have the following calculation

$d_{min} = \min_{c' \neq c} D_{P^{NM} \cup P^M}(c, c') = -4|M| + |M| = -3|M|$.

$D_{P^{NM} \cup P^M}(u, v') = -4|M| + (|M| - 1) - 1 = -3|M| - 2 < -3|M| = d_{min}$.

$D_{P^{NM} \cup P^M}(v, u') = -4|M| + 1 - (|M| + 1) = -5|M| + 2 < -3|M| = d_{min}$.

For any $t \in \mathcal{C} \setminus \{c, u, v\}$, there exists $s \in \mathcal{C} \setminus \{c\}$ such that $(s, t) \in E$ and $D_{P^M}(t, s) = -|M|$, which means that $D_{P^{NM} \cup P^M}(t, s) = -2|M| - 2 - |M| = -3|M| - 2 < -3|M| = d_{min}$.

Hence $Maximin(P^{NM} \cup P^M) = \{c\}$.

Next, we prove that if there exists a profile P^M for the manipulators such that $Maximin(P^{NM} \cup P^M) = \{c\}$, then there exist two vertex disjoint paths from u to u' and from v to v'. We define a function $f : V \to V$ such that $D_{P^{NM} \cup P^M}(t, f(t)) < -3|M|$. We note that since $Maximin(P^{NM} \cup P^M) = \{c\}$, for any $t \neq c$, there must exist s such that $D_{P^{NM} \cup P^M}(t, s) < -3|M|$, and s must be a parent of t in G'. If there exists more than one such s, define $f(t)$ to be any one of them. It follows that if $(t, f(t))$ is neither (u, v') or (v, u'), then $(f(t), t) \in E$ and $D_{P^M}(t, f(t)) = -|M|$, which means that $f(t) \succ t$ in each vote of P^M; otherwise, if $(t, f(t))$ is (u, v') or (v, u'), then $D_{P^M}(t, f(t)) \leq |M| - 2$, which means that $f(t) \succ t$ in at least one vote of P^M. There must exist $l_1 < l_2 \leq m$ such that $f^{l_1}(u) = f^{l_2}(u)$. That is, $f^{l_1}(u), f^{l_1+1}(u), \ldots, f^{l_2-1}(u), f^{l_2}(u)$ is a cycle in G'. We assume that for any $l_1 \leq l_1' < l_2' < l_2$, $f^{l_1'}(u) \neq f^{l_2'}(u)$. Now we claim that (v', u) and (u', v) must be both in the cycle, because

1. if neither of them is in the cycle, then in each vote of P^M, we must have $f^{l_2}(u) \succ f^{l_2-1}(u) \succ f^{l_1}(u) = f^{l_2}(u)$, which contradicts the assumption that each vote is a linear order;
2. if exactly one of them is in the cycle—without loss of generality, $f^{l_1}(u) = v, f^{l_1+1}(u) = u'$—then in at least one of the votes of P^M, we must have $f^{l_2}(u) \succ f^{l_2-1}(u) \succ \ldots \succ f^{l_1}(u) = f^{l_2}(u)$, which contradicts the assumption that each vote is a linear order.

Now, without loss of generality, let us assume that $f^{l_1}(u) = u, f^{l_1+1}(u) = v'$, $f^{l_3}(u) = v, f^{l_3+1}(u) = u'$, where $l_3 \leq l_2 - 2$. We immediately obtain two vertex disjoint paths $u = f^{l_1}(u) = f^{l_2}(u) \to f^{l_2-1}(u) \to \ldots \to f^{l_3+1}(u) = u'$ and $v = f^{l_3}(u) \to f^{l_3-1}(u) \to \ldots \to f^{l_1+1}(u) = v'$. Therefore, UCMU under maximin is NP-complete.

For UCMC, we use almost the same reduction, except we modify it as follows:

2'. Let $D_{P^{NM}}(u, v') = D_{P^{NM}}(v, u') = -4|M| + 2$.

3'. For any $(s, t) \in E$ such that $D_{P^{NM}}(t, s)$ is not defined above, we let $D_{P^{NM}}(t, s) = -2|M|$.

□

4 Ranked pairs

In this section, we prove that the UCMU and UCMC problems under ranked pairs are NP-complete (even for a single manipulator) by giving a reduction from 3SAT.

Definition 5 *The* 3SAT *problem is: Given a set of variables* $X = \{x_1, \ldots, x_q\}$ *and a formula* $Q = Q_1 \wedge \ldots \wedge Q_t$ *such that*

1. *for all* $1 \leq i \leq t$, $Q_i = l_{i,1} \vee l_{i,2} \vee l_{i,3}$, *and*
2. *for all* $1 \leq i \leq t$ *and* $1 \leq j \leq 3$, $l_{i,j}$ *is either a variable* x_k, *or the negation of a variable* $\neg x_k$,

we are asked whether the variables can be set to true or false so that Q *is true.*

Theorem 2 *The UCMU and UCMC problems under ranked pairs are NP-complete, even when there is only one manipulator.*

Proof of Theorem 2: It is easy to verify that the UCMU and UCMC problems under ranked pairs are in NP. We first prove that UCMU is NP-complete. Given an instance of 3SAT, we construct a UCMU instance as follows. Without loss of generality, we assume that for any variable x, x and $\neg x$ appears in at least one clause, and none of the clauses contain both x and $\neg x$.
Set of alternatives: $\mathcal{C} = \{c, Q_1, \ldots, Q_t, Q'_1, \ldots, Q'_t\} \bigcup \{x_1, \ldots, x_q, \neg x_1, \ldots, \neg x_q\}$
$\bigcup \{Q_{l_{1,1}}, Q_{l_{1,2}}, Q_{l_{1,3}}, \ldots, Q_{l_{t,1}}, Q_{l_{t,2}}, Q_{l_{t,3}}\} \bigcup \{Q_{\neg l_{1,1}}, Q_{\neg l_{1,2}}, Q_{\neg l_{1,3}}, \ldots, Q_{\neg l_{t,1}}, Q_{\neg l_{t,2}}, Q_{\neg l_{t,3}}\}$.
Alternative preferred by the manipulator: c.
Number of unweighted manipulators: $|M| = 1$.
Non-manipulators' profile: P^{NM} satisfying the following conditions.

1. For any $i \leq t$, $D_{P^{NM}}(c, Q_i) = 30$,$D_{P^{NM}}(Q'_i, c) = 20$; for any $x \in \mathcal{C} \setminus \{Q_i, Q'_i : 1 \leq i \leq t\}$, $D_{P^{NM}}(c, x) = 10$.
2. For any $j \leq q$, $D_{P^{NM}}(x_j, \neg x_j) = 20$.
3. For any $i \leq t, j \leq 3$, if $l_{i,j} = x_k$, then $D_{P^{NM}}(Q_i, Q^i_{x_k}) = 30$, $D_{P^{NM}}(Q^i_{x_k}, x_k) = 30$, $D_{P^{NM}}(\neg x_k, Q^i_{\neg x_k}) = 30$, $D_{P^{NM}}(Q^i_{\neg x_k}, Q'_i) = 30$; if $l_{i,j} = \neg x_k$, then $D_{P^{NM}}(Q_i, Q^i_{\neg x_k}) = 30$, $D_{P^{NM}}(Q^i_{x_k}, x_k) = 30$, $D_{P^{NM}}(\neg x_k, Q^i_{\neg x_k}) = 30$, $D_{P^{NM}}(Q^i_{x_k}, Q'_i) = 30$, $D_{P^{NM}}(Q^i_{\neg x_k}, Q^i_{x_k}) = 20$.
4. For any $x, y \in \mathcal{C}$, if $D_{P^{NM}}(x, y)$ is not defined in the above steps, then $D_{P^{NM}}(x, y) = 0$.

For example, when $Q_1 = x_1 \vee \neg x_2 \vee x_3$, $D_{P^{NM}}$ is illustrated in Figure 1.

The existence of such a P^{NM} is guaranteed by Lemma 1, and the size of P^{NM} is in polynomial in t and q.

First, we prove that if there exists an assignment v of truth values to X so that Q is satisfied, then there exists a vote R_M for the manipulator such that $RP(P^{NM} \cup \{R_M\}) = \{c\}$. We construct R_M as follows.

- Let c be on the top of R_M.
- For any $k \leq q$, if $v(x_k) = \top$ (that is, x_k is true), then $x_k \succ_{R_M} \neg x_k$, and for any $i \leq t, j \leq 3$ such that $l_{i,j} = \neg x_k$, let $Q^i_{x_k} \succ_{R_M} Q^i_{\neg x_k}$.
- For any $k \leq q$, if $v(x_k) = \bot$ (that is, x_k is false), then $\neg x_k \succ_{R_M} x_k$, and for any $i \leq t, j \leq 3$ such that $l_{i,j} = \neg x_k$, let $Q^i_{\neg x_k} \succ_{R_M} Q^i_{x_k}$.
- The remaining pairs of alternatives are ranked arbitrarily.

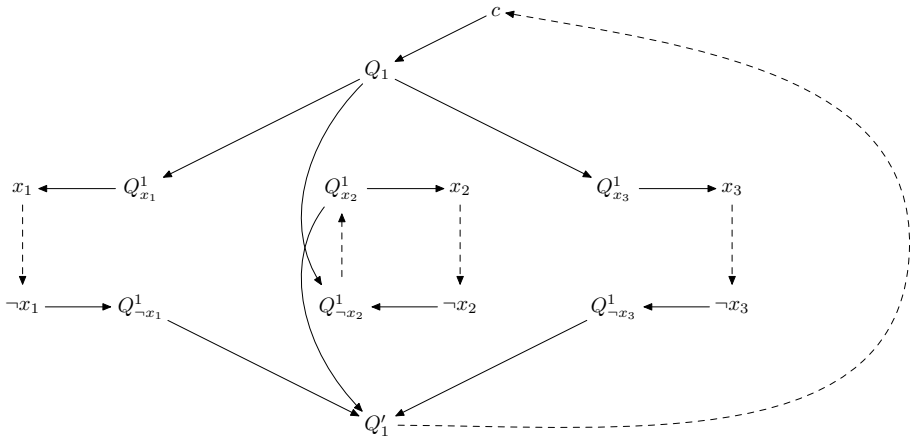


Figure 1: For any vertices v_1, v_2, if there is a solid edge from v_1 to v_2, then $D_{P^{NM}}(v_1, v_2) = 30$; if there is a dashed edge from v_1 to v_2, then $D_{P^{NM}}(v_1, v_2) = 20$; if there is no edge between v_1 and v_2 and $v_1 \neq c$, $v_2 \neq c$, then $D_{P^{NM}}(v_1, v_2) = 0$; for any x such that there is no edge between c and x, $D_{P^{NM}}(c, x) = 10$.

If $x_k = \top$, then $D_{P^{NM}\cup\{R_M\}}(x_k, \neg x_k) = 21$, and for any $i \leq t, j \leq 3$ such that $l_{i,j} = \neg x_k$, $D_{P^{NM}\cup\{R_M\}}(Q^i_{\neg x_k}, Q^i_{x_k}) = 19$. It follows that no matter how ties are broken when applying ranked pairs to $P^{NM} \cup \{R_M\}$, if $x_k = \top$, then $x_k \succ \neg x_k$ in the final ranking. This is because for any $l_{i,j} = \neg x_k$, $D_{P^{NM}\cup\{R_M\}}(Q^i_{\neg x_k}, Q^i_{x_k}) = 19 < 21 = D_{P^{NM}\cup\{R_M\}}(x_k, \neg x_k)$, which means that before trying to fix $x_k \succ \neg x_k$, there is no directed path from $\neg x_k$ to x_k.

Similarly if $x_k = \bot$, then $D_{P^{NM}\cup\{R_M\}}(x_k, \neg x_k) = 19$, and for any $i \leq t, j \leq 3$ such that $l_{i,j} = \neg x_k$, $D_{P^{NM}\cup\{R_M\}}(Q^i_{\neg x_k}, Q^i_{x_k}) = 21$. It follows that if $x_k = \bot$, then $\neg x_k \succ x_k$, and for any $i \leq t, j \leq 3$ such that $l_{i,j} = \neg x_k$, $Q^i_{\neg x_k} \succ Q^i_{x_k}$ in the final ranking. This is because $Q^i_{\neg x_k} \succ Q^i_{x_k}$ will be fixed before $x_k \succ \neg x_k$.

Because Q is satisfied under v, for each clause Q_i, at least one of its three literals is true under v. Without loss of generality, we assume $v(l_{i,1}) = \top$. If $l_{i,1} = x_k$, then before trying to add $Q'_i \succ c$, the directed path $c \to Q_i \to Q_{x_k} \to x_k \to \neg x_k \to Q_{\neg x_k} \to Q'_i$ has already been fixed. Therefore, $c \succ Q'_i$ in the final ranking, which means that for any alternatives x in $\mathcal{C} \setminus \{c, Q_1, \ldots, Q_t, Q'_1, \ldots, Q'_t\}$, $c \succ x$ in the final ranking because $D_{P^{NM}\cup\{R_M\}}(c, x) > 0$. Hence, c is the unique winner of $P^{NM} \cup \{R_M\}$ under ranked pairs.

Next, we prove that if there exists a vote R_M for the manipulator such that $RP(P^{NM}\cup\{R_M\}) = \{c\}$, then there exists an assignment v of truth values to X such that Q is satisfied. We construct the assignment v so that $v(x_k) = \top$ if and only if $x_k \succ_{R_M} \neg x_k$, and $v(x_k) = \bot$ if and only if $\neg x_k \succ_{R_M} x_k$. We claim that $v(Q) = \top$. If, on the contrary, $v(Q) = \bot$, then there exists a clause (Q_1, without loss of generality) such that $v(Q_1) = \bot$. We now construct a way to fix the pairwise rankings such that c is not the winner under ranked pairs, as follows. For any $j \leq 3$, if there exists $k \leq q$ such that $l_{i,j} = \neg x_k$, then $x_k \succ_{R_M} \neg x_k$ because $v(\neg x_k) = \bot$. Therefore, $D_{P^{NM}\cup R_M}(x_k, \neg x_k) = 21$. Then, after trying to add all pairs $x \succ x'$ such that $D_{P^{NM}\cup R_M}(x, x') > 21$ (that is, all solid directed edges in Figure 1), it follows that $x_k \succ \neg x_k$ can be added to the final ranking. We choose to add $x_k \succ \neg x_k$ first, which means that $Q^1_{x_k} \succ Q^1_{\neg x_k}$ in the final ranking (otherwise, we have $Q^1_{\neg x_k} \succ Q^1_{x_k} \succ x_k \succ \neg x_k \succ Q^1_{\neg x_k}$, which is a contradiction).

For any $j \leq 3$, if there exists $k \leq q$ such that $l_{i,j} = x_k$, then $\neg x_k \succ_{R_M} x_k$ because $v(x_k) = \bot$. Therefore, $D_{P^{NM}\cup R_M}(x_k, \neg x_k) = 19$. We note that after trying to add all pairs $x \succ x'$ such that $D_{P^{NM}\cup R_M}(x, x') > 19$, $Q^1_{x_k} \not\succ Q^1_{\neg x_k}$. We recall that for any $j \leq 3$, if there exists $k \leq q$ such that $l_{i,j} = \neg x_k$, then $Q^1_{\neg x_k} \not\succ Q^1_{x_k}$. Hence, it follows that $Q'_1 \succ c$ is consistent with all pairwise rankings

added so far. Then, since $D_{P^{NM} \cup R_M}(Q'_1, c) \geq 19$, if $Q'_1 \succ c$ has not been added, we choose to add it first of all pairwise rankings of alternatives $x \succ x'$ such that $D_{P^{NM} \cup R_M}(x, x') = 19$, which means that $Q'_1 \succ c$ in the final ranking—in other words, c is not at the top in the final ranking. Therefore, c is not the unique winner, which contradicts the assumption that $RP(P^{NM} \cup \{R_M\}) = \{c\}$.

For UCMC, we modify the reduction as follows: we let P^{NM} be such that for any $i \leq t$, $D_{P^{NM}}(Q'_i, c) = 22$, and for any $j \leq q$, $D_{P^{NM}}(x_j, \neg x_j) = 22$. □

Similarly, we can prove that when $|M|$ is a constant greater than one, UCMU and UCMC under ranked pairs remain NP-complete.

Theorem 3 *The UCMU and UCMC problems under ranked pairs are NP-complete, even when the number of manipulators is fixed to some constant* $|M| > 1$.

Proof of Theorem 3: We prove UCMU is NP-complete. The proof is similar to that of Theorem 2. We let P^{NM} satisfy the following conditions.

1. For any $i \leq t$, $D_{P^{NM}}(c, Q_i) = 30|M|$,$D_{P^{NM}}(Q'_i, c) = 22|M| - 2$; for any $x \in \mathcal{C} \setminus \{Q_i, Q'_i : 1 \leq i \leq t\}$, $D_{P^{NM}}(c, x) = 10|M|$.

2. For any $j \leq q$, $D_{P^{NM}}(x_j, \neg x_j) = 22|M| - 2$.

3. For any $i \leq t, j \leq 3$, if $l_{i,j} = x_k$, then $D_{P^{NM}}(Q_i, Q^i_{x_k}) = 30|M|$, $D_{P^{NM}}(Q^i_{x_k}, x_k) = 30|M|$, $D_{P^{NM}}(\neg x_k, Q^i_{\neg x_k}) = 30|M|$, $D_{P^{NM}}(Q^i_{\neg x_k}, Q'_i) = 30|M|$; if $l_{i,j} = \neg x_k$, then $D_{P^{NM}}(Q_i, Q^i_{\neg x_k}) = 30|M|$, $D_{P^{NM}}(Q^i_{x_k}, x_k) = 30|M|$, $D_{P^{NM}}(\neg x_k, Q^i_{\neg x_k}) = 30|M|$, $D_{P^{NM}}(Q^i_{x_k}, Q'_i) = 30|M|$, $D_{P^{NM}}(Q^i_{\neg x_k}, Q^i_{x_k}) = 20|M|$.

4. For any $x, y \in \mathcal{C}$, if $D_{P^{NM}}(x, y)$ is not defined in the above steps, then $D_{P^{NM}}(x, y) = 0$.

First, if there exists an assignment v of truth values to X so that Q is satisfied, then we let R_M be defined as in the proof for Theorem 2. It follows that $RP(P^{NM} \cup \{|M|R_M\}) = \{c\}$ (all the manipulators can vote R_M).

Next, if there exists a profile P^M for the manipulators such that $RP(P^{NM} \cup P^M) = \{c\}$, then we construct the assignment v so that $v(x_k) = \top$ if $x_k \succ_V \neg x_k$ for all $V \in P^M$, and $v(x_k) = \bot$ if $\neg x_k \succ_V x_k$ for all $V \in P^M$; the values of all the other variables are assigned arbitrarily. Then by similar reasoning as in the proof for Theorem 2, we know that Q is satisfied under v.

For UCMC, the proof is similar (by slightly modifying the $D_{P^{NM}}$ as we did in the proof of Theorem 2). □

5 Bucklin

In this section, we present a polynomial-time algorithm for the UCMU problem under Bucklin (a polynomial-time algorithm for the UCMC problem under Bucklin can be obtained similarly). For any alternative x, any natural number d, and any profile P, let $B(x, d, P)$ denote the number of times that x is ranked among the top d alternatives in P. The idea behind the algorithm is as follows. Let d_{min} be the minimal depth so that c is ranked among the top d_{min} alternatives in more than half of the votes (when all of the manipulators rank c first). Then, we check if there is a way to assign the manipulators' votes so that none of the other alternatives is ranked among the top d_{min} alternatives in more than half of the votes.

Algorithm 1
Input: *A UCM instance* $(Bucklin, P^{NM}, c, M)$, $C = \{c, c_1, \ldots, c_{m-1}\}$.

1. *Calculate the minimal depth* d_{min} *such that* $B(c, d_{min}, P^{NM}) + |M| > \frac{1}{2}(|NM| + |M|)$.

2. *If there exists $c' \in C$, $c' \neq c$ such that $B(c', d_{min}, P^{NM}) > \frac{1}{2}(|NM| + |M|)$, then output that there is no successful manipulation. Otherwise, for any $c' \in C$, $c' \neq c$, let $d_{c'} = \lfloor \frac{1}{2}(|NM| + |M|) \rfloor - B(c', d_{min}, P^{NM})$, $k_{c'} = \begin{cases} |M| & \text{if } d_{c'} \geq |M| \\ d_{c'} & \text{otherwise} \end{cases}$.*

3. *If $\sum_{c' \neq c} k_{c'} < (d_{min} - 1)|M|$, then output that there is no successful manipulation.*

4. *Let $j = 1$, $t = 1$, and for any $l \leq |M|$, let R_l rank c at the top. Repeat Step 4a $m - 1$ times:*

4a. *If $k_{c_t} > 0$, then c_t is ranked in the next position (lower than the candidates that have already been ranked in previous steps) in $R_{\mod (j-1,|M|)+1}, R_{\mod (j,|M|)+1}, \ldots, R_{\mod (j+k_{c_t}-2,|M|)+1}$, respectively, where for any natural number a, b, $\mod (a, b)$ is the common residue of $a (\mod b)$. Let $j \leftarrow \mod (j + k_{c_t} - 1, |M|) + 1$, $t \leftarrow t + 1$.*

5. *For any $s \leq |M|$, complete R_s arbitrarily. Output $P^M = (R_1, \ldots, R_{|M|})$.*

Claim 1 *Algorithm 1 correctly solves the UCMU problem. It runs in time $O(m|NM| + |NM||M| + |M|m)$.*

Proof of Claim 1: Let us first consider the case where Algorithm 1 outputs that there is no successful manipulation. There are two cases.

1. There exists $c' \in C$, $c' \neq c$ such that $B(c', d_{min}, P^{NM}) > \frac{1}{2}(|NM| + |M|)$.
2. $\sum_{c' \neq c} k_{c'} < (d_{min} - 1)|M|$. In this case, for any P^M, there exists $c' \neq c$ such that $|M| \geq B(c', d_{min}, P^M) > k_{c'}$, which means that $B(c', d_{min}, P^{NM} \cup P^M) > \frac{1}{2}(|NM| + |M|)$.

In both cases, more than half of the voters rank c' among the top d_{min} alternatives. Therefore, c cannot be the unique winner.

Now let us consider the case where Algorithm 1 outputs some P^M. In this case, for any $t \leq m - 1$, $B(c_t, d_{min}, P^M) \leq k_{c_t}$. Therefore, for any $t \leq m - 1$, $B(c_t, d_{min}, P^{NM} \cup P^M) \leq B(c_t, d_{min}, P^{NM}) + k_{c_t} \leq \frac{1}{2}(|NM| + |M|)$, which means that $Bucklin(P^{NM} \cup P^M) = \{c\}$.

Step 1 runs in time $O(m|NM|)$, Step 2 runs in time $O(|M||NM|)$, Step 3 runs in time $O(|M|)$, and Step 4 and Step 5 run in time $O(m|M|)$. Therefore, Algorithm 1 runs in time $O(m|NM| + |NM||M| + |M|m)$. □

6 Discussion

Number of manipulators	1	constant
Copeland (specific tie-breaking)	P [2]	NP-hard [8]
STV	NP-hard [1]	NP-hard [1]
Veto	P [20]	P [20]
Plurality with Runoff	P [20]	P [20]
Cup	P [6]	P [6]
Maximin	P [2]	**NP-hard**
Ranked pairs	**NP-hard**	**NP-hard**
Bucklin	**P**	**P**
Borda	P [2]	?

Table 1: Complexity of UCM under prominent voting rules. Boldface results appear in this paper.

In this paper, we studied the computational complexity of unweighted coalitional manipulation under the maximin, ranked pairs, and Bucklin rules. The UCM problem is NP-complete under the maximin rule for any fixed number (at least two) of manipulators. The UCM problem is also NP-complete under the ranked pairs rule; in this case, the hardness holds even if there is only a single manipulator, similarly to the second-order Copeland and STV rules. We gave a polynomial-time algorithm for the UCM problem under the Bucklin rule. Table 1 summarizes our results, and puts them in the context of previous results on the UCM problem.

It should be noted that all of these hardness results, as well as the ones mentioned in the introduction, are *worst-case* results. Hence, there may still be an efficient algorithm that can find a beneficial manipulation for *most* instances. Indeed, several recent results suggest that finding manipulations is usually easy. Procaccia and Rosenschein have shown that, when the number of alternatives is a constant, manipulation of positional scoring rules is easy even with respect to "junta" distributions, which arguably focus on hard instances [15]. Conitzer and Sandholm have given some sufficient conditions under which manipulation is easy and argue that these conditions are usually satisfied in practice [5]. Zuckerman et al. have given manipulation algorithms with the property that if they fail to find a manipulation when one exists, then, if the manipulators are given some additional vote weights, the algorithm will succeed [20]. The asymptotic probability of manipulability has also been characterized (except for knife-edge cases) for a very general class of voting rules [18] (building on earlier work [14]). In a similar spirit, several quantitative versions of the Gibbard-Satterthwaite theorem have recently been proved [9, 19]. One weakness of all of these results (except [20]) is that they make assumptions about the distribution of instances. In this paper, we have focused on the worst-case framework, which does not suffer from this weakness. This does mean that when we show that manipulation is hard, it may still be the case that it is usually easy.

There are many interesting problems left for future research. For example, settling the complexity of UCM under positional scoring rules such as Borda is a challenging open problem.

Acknowledgments

We thank anonymous reviewers for helpful comments and suggestions. This work is supported in part by the United States-Israel Binational Science Foundation under grant 2006-216. Lirong Xia is supported by a James B. Duke Fellowship, Vincent Conitzer is supported by an Alfred P. Sloan Research Fellowship, and Ariel Procaccia is supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities. Xia and Conitzer are also supported by the NSF under award number IIS-0812113.

References

[1] John Bartholdi, III and James Orlin. Single transferable vote resists strategic voting. *Social Choice and Welfare*, 8(4):341–354, 1991.

[2] John Bartholdi, III, Craig Tovey, and Michael Trick. The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6(3):227–241, 1989.

[3] Vincent Conitzer and Tuomas Sandholm. Universal voting protocol tweaks to make manipulation hard. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 781–788, Acapulco, Mexico, 2003.

[4] Vincent Conitzer and Tuomas Sandholm. Common voting rules as maximum likelihood estimators. In *Proceedings of the 21st Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 145–152, Edinburgh, UK, 2005.

[5] Vincent Conitzer and Tuomas Sandholm. Nonexistence of voting rules that are usually hard to manipulate. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, Boston, MA, 2006.

[6] Vincent Conitzer, Tuomas Sandholm, and Jérôme Lang. When are elections with few candidates hard to manipulate? *Journal of the ACM*, 54(3):Article 14, 1–33, 2007. Early versions in AAAI-02 and TARK-03.

[7] Edith Elkind and Helger Lipmaa. Hybrid voting protocols and hardness of manipulation. In *Annual International Symposium on Algorithms and Computation (ISAAC)*, 2005.

[8] Piotr Faliszewski, Edith Hemaspaandra, and Henning Schnoor. Copeland voting: Ties matter. In *To appear in Proceedings of AAMAS'08*, 2008.

[9] Ehud Friedgut, Gil Kalai, and Noam Nisan. Elections can be manipulated often. In *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2008.

[10] Allan Gibbard. Manipulation of voting schemes: a general result. *Econometrica*, 41:587–602, 1973.

[11] Edith Hemaspaandra and Lane A. Hemaspaandra. Dichotomy for voting systems. *Journal of Computer and System Sciences*, 73(1):73–83, 2007.

[12] Andrea S. LaPaugh and Ronald L. Rivest. The subgraph homeomorphism problem. In *Proceedings of the tenth annual ACM symposium on Theory of computing (STOC)*, pages 40–50, 1978.

[13] David C. McGarvey. A theorem on the construction of voting paradoxes. *Econometrica*, 21(4):608–610, 1953.

[14] Ariel D. Procaccia and Jeffrey S. Rosenschein. Average-case tractability of manipulation in voting via the fraction of manipulators. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, Honolulu, HI, USA, 2007.

[15] Ariel D. Procaccia and Jeffrey S. Rosenschein. Junta distributions and the average-case complexity of manipulating elections. *Journal of Artificial Intelligence Research (JAIR)*, 28:157–181, 2007.

[16] Mark Satterthwaite. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10:187–217, 1975.

[17] T. N. Tideman. Independence of clones as a criterion for voting rules. *Social Choice and Welfare*, 4(3):185–206, 1987.

[18] Lirong Xia and Vincent Conitzer. Generalized scoring rules and the frequency of coalitional manipulability. In *Proceedings of the Ninth ACM Conference on Electronic Commerce (EC)*, 2008.

[19] Lirong Xia and Vincent Conitzer. A sufficient condition for voting rules to be frequently manipulable. In *Proceedings of the Ninth ACM Conference on Electronic Commerce (EC)*, 2008.

[20] Michael Zuckerman, Ariel D. Procaccia, and Jeffrey S. Rosenschein. Algorithms for the coalitional manipulation problem. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2008.

Lirong Xia
Department of Computer Science
Duke University
Durham, NC 27708, USA
Email: lxia@cs.duke.edu

Vincent Conitzer
Department of Computer Science
Duke University
Durham, NC 27708, USA
Email: conitzer@cs.duke.edu

Ariel Procaccia
School of Engineering and Computer Science
The Hebrew University of Jerusalem
Jerusalem, Israel
Email: arielpro@cs.huji.ac.il

Jeffrey S. Rosenschein
School of Engineering and Computer Science
The Hebrew University of Jerusalem
Jerusalem, Israel
Email: jeff@cs.huji.ac.il