Group Manipulation in Judgment Aggregation

Sirin Botan
ILLC, University of Amsterdam
botan.sirin@gmail.com

Arianna Novaro
ILLC, University of Amsterdam
arianna.novaro@gmail.com

Ulle Endriss
ILLC, University of Amsterdam
ulle.endriss@uva.nl

ABSTRACT

We introduce the concept of group manipulation into the study of judgment aggregation and investigate the circumstances under which an aggregation rule may be subject to strategic misrepresentation of judgments by a group of agents. Our focus is on neutral aggregation rules, which treat all propositions to be judged symmetrically, and we assume that agents strategise to minimise the number of propositions on which they disagree with the outcome of a rule. We find that strategic manipulation by groups of two agents can be ruled out for the independent and monotonic aggregation rules. This family of rules, which is precisely the family of rules for which manipulation by a single agent can be ruled out, includes the widely used uniform quota rules. When three or more agents may coordinate their manipulation, on the other hand, essentially all attractive rules are susceptible to strategic manipulation. However, we are able to recover the family of independent and monotonic rules as being immune to manipulation, if we add the assumption that the members of a group of manipulating agents fear that the others might opt out of the jointly agreed plan.

Keywords

Collective Decision Making, Social Choice Theory

1. INTRODUCTION

Judgment aggregation [16] is a formal framework for integrating the views of several autonomous agents into a single collective view. Originating in Philosophy and Economics, and inspired by the *doctrinal paradox* observed in Legal Theory [15], it has recently received increased attention in Computer Science and Artificial Intelligence [7], where the related framework of belief revision had been studied for some time already [10, 14]. One reason for this interest is its great potential for applications, e.g., for argumentation in multiagent systems [20] or for crowdsourcing [19].

Agents will sometimes want to manipulate the process of aggregation by misreporting their true judgments. This, the analysis of strategic behaviour, is central to the closely related study of voting and preference aggregation, in both Economics [1] and Computer Science [3], but it has received very little attention in judgment aggregation to date. The reason is that here, unlike for preference aggregation, there is no single most natural way of modelling the incentives of a strategic agent. One concrete proposal for modelling such incentives is due to Dietrich and List [5], who initiated the study of strategic manipulation in judgment aggregation. They proposed a simple manner in which to induce preferences from an agent's judgment set, namely to assume that an agent's preferences only depend on the number of propositions for which their own judgment agrees with that of the rule. This initial contribution on strategic manipulation in judgment aggregation, focussing on axiomatic characterisation results, was later followed up by work on the computational complexity of manipulation by Endriss et al. [8] and Baumeister et al. [2].

All of these contributions have dealt with manipulation by a single agent at a time, rather than with the coordinated manipulation by a group. The same holds for related work on strategic manipulation in belief merging [9]. In this paper, we define a suitable notion of group manipulation for judgment aggregation and characterise the class of rules that are immune to this kind of interference. In doing so, we focus on neutral aggregation rules, i.e., rules that treat all propositions to be judged symmetrically. While our main results are axiomatic and apply to large families of rules that meet certain desirable axiomatic properties (notably neutrality, a closely related property called unbiasedness we introduce here, independence, and monotonicity), we pay special attention to a particularly natural family of concrete aggregation rules often used in practice, namely the *uniform quota rules*, and discuss the impact of our results on these rules in some detail.

The remainder of this paper is organised as follows. In Section 2 we recall relevant concepts from the judgment aggregation literature and fix our notation and terminology. In Section 3 we discuss previous work on strategic manipulation by a single agent and prove a new characterisation result for aggregation rules that are immune to single-agent manipulation. This result then serves as a base line for our work on group manipulation. Our main results can be found in Section 4, where we define the notion of group manipulation and characterise the family of aggregation rules that are immune to it, showing that essentially all reasonable rules are susceptible to this form of manipulation. This is true, in particular, for all (nontrivial) uniform quota rules, even though these rules are known to be single-agent strategyproof. Following this negative result, in Section 5 we consider a variant of our main notion of group manipula-

Appears in: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016), J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.), May 9–13, 2016, Singapore.*
Copyright © 2016, International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

tion, under which individual agents may opt out of a coalition of manipulators. We show that under this alternative definition we can obtain more positive results, and that the difference between immunity against single-agent and group manipulation disappears again. Section 6 concludes.

2. PRELIMINARIES

In this section, we introduce the relevant notation and terminology to speak about judgment aggregation (JA). Specifically, we shall work with the standard framework of JA [7, 13, 17], going back to the work of List and Pettit [16]. We also introduce a number of novel concepts we require for our purposes that may be of independent interest, namely the axiom of *unbiasedness* and the notions of *flipping* a formula in a judgment set and of *restricting* an aggregation rule.

2.1 Basic Notation and Terminology

An *agenda* is a finite set of formulas of propositional logic of the form $\Phi = \Phi^+ \cup \{\neg\varphi \mid \varphi \in \Phi^+\}$, such that Φ^+, the *pre-agenda*, only contains formulas that are not negated. We call Φ *atomic* in case Φ^+ only contains atomic propositions.[1] A *judgment set* for Φ is a subset $J \subseteq \Phi$. We call J *complete* if $\varphi \in J$ or $\neg\varphi \in J$ for all $\varphi \in \Phi^+$; we call J *complement-free* if $\varphi \notin J$ or $\neg\varphi \notin J$ for all $\varphi \in \Phi^+$; and we call J *consistent* if it is logically consistent. $\mathcal{J}(\Phi)$ denotes the set of all complete and consistent judgment sets over Φ.

Let $\mathcal{N} = \{1, \ldots, n\}$ be a finite set of *agents*. A *profile* $\boldsymbol{J} = (J_1, \ldots, J_n)$ is a vector of judgment sets, one for each agent. Agent i is said to *accept* formula $\varphi \in \Phi$ if $\varphi \in J_i$; otherwise she *rejects* φ. We write $N_\varphi^{\boldsymbol{J}} = \{i \in \mathcal{N} \mid \varphi \in J_i\}$ for the *coalition of supporters* of φ in profile $\boldsymbol{J}$, i.e., for the set of agents who accept φ in $\boldsymbol{J}$. We furthermore write $(\boldsymbol{J}_{-i}, J_i')$ for the profile that is like $\boldsymbol{J}$, except that J_i has been replaced by J_i'. We say that profiles $\boldsymbol{J}$ and $\boldsymbol{J'}$ are C-*variants*, for some coalition $C \subseteq \mathcal{N}$, if $J_i = J_i'$ for all $i \in \mathcal{N} \setminus C$. A profile is called *unanimous* if it is of the form $\boldsymbol{J} = (J, \ldots, J)$.

For $\varphi \in \Phi^+$, we denote with $J^{\rightleftarrows\varphi}$ the result of *flipping* φ in judgment set J (i.e., of replacing φ by $\neg\varphi$ or $\neg\varphi$ by φ). Analogously, for $S \subseteq \Phi^+$, we denote with $J^{\rightleftarrows S}$ the result of flipping all the formulas in S in judgment set J. For example, $\{p, \neg q, r\}^{\rightleftarrows\{p,q\}} = \{\neg p, q, r\}$. Note that this flipping operation is cardinality-preserving, i.e., we always have $|J| = |J^{\rightleftarrows S}|$, for any J and S. Finally, we use $\boldsymbol{J}^{\rightleftarrows S}$ to denote the result of flipping the formulas in S in all the judgment sets in a given profile $\boldsymbol{J}$.

2.2 Aggregation Rules

An *aggregation rule* is a function $F : \mathcal{J}(\Phi)^n \to 2^\Phi$ that assigns a collective judgment set $F(\boldsymbol{J}) \subseteq \Phi$ to every complete and consistent profile (2^Φ denotes the powerset of Φ). If $\varphi \in F(\boldsymbol{J})$, then that means that the group accepts φ.

An example for an aggregation rule is the (weak) *majority rule*, which accepts a formula if and only if at least half of the agents do. A *uniform quota rule* F_q is an aggregation rule induced by a number $q \in \{0, 1, \ldots, n+1\}$ such that:

$$F_q(\boldsymbol{J}) \quad = \quad \{\varphi \in \Phi \mid \#N_\varphi^{\boldsymbol{J}} \geqslant q\}$$

Thus, we obtain the majority rule for $q = \lceil \frac{n}{2} \rceil$. The *unanimity rule* is the uniform quota rule with quota $q = n$ (requiring all agents to accept a formula for it to get accepted) and the *nomination rule* is the uniform quota rule with $q = 1$ (requiring acceptance by at least one agent).

A *constant rule* is a rule that returns the same fixed judgment set for every possible profile. Two examples for constant rules are the *trivial uniform quota rules* with quotas $q = 0$ and $q = n + 1$, respectively, which always accept or reject all formulas, respectively. The *dictatorship* of agent $i \in \mathcal{N}$ is the rule $F^i : \boldsymbol{J} \mapsto J_i$, returning the judgment set of i, whatever the judgments of the others.

We define the *restriction* of an aggregation rule F to a coalition $C \subseteq \mathcal{N}$ and a subset $\Psi^+ \subseteq \Phi^+$ of the pre-agenda for a given profile $\boldsymbol{J} \in \mathcal{J}(\Phi)^n$ as the following aggregation rule F', defined for the agents in C and the agenda $\Psi := \Psi^+ \cup \{\neg\psi \mid \psi \in \Psi^+\}$. F' receives a profile in $\mathcal{J}(\Psi)^{|C|}$ as input. We then amend this small profile to obtain a large profile in $\mathcal{J}(\Phi)^n$ by fixing the judgment sets of agents not in C as in $\boldsymbol{J}$ and by fixing the judgments of agents in C on agenda formulas not in Ψ also as in $\boldsymbol{J}$. Finally, we compute the output of F for the large profile and return the judgments as far as the agenda formulas in Ψ are concerned as the output of F'.[2] Note that for every C and Ψ^+, there are several restrictions of F (namely one for every $\boldsymbol{J}$).

2.3 Axioms

Any given rule may or may not satisfy a number of normatively desirable properties, usually referred to as *axioms*. Several such axioms will play a role in this paper.

First, F is *independent* if $N_\varphi^{\boldsymbol{J}} = N_\varphi^{\boldsymbol{J'}}$ implies $\varphi \in F(\boldsymbol{J}) \Leftrightarrow \varphi \in F(\boldsymbol{J'})$, i.e., if acceptance of a formula only depends on which agents accept that same formula. Second, F is *monotonic* if $\varphi \in J_i' \setminus J_i$ for some $i \in \mathcal{N}$ implies $\varphi \in F(\boldsymbol{J}) \Rightarrow \varphi \in F(\boldsymbol{J}_{-i}, J_i')$, i.e., if additional support for an accepted formula never causes that formula to be rejected. The following simple result provides a convenient characterisation of rules that are both independent and monotonic.

LEMMA 1. *An aggregation rule F is both independent and monotonic if and only if, for all profiles $\boldsymbol{J}, \boldsymbol{J'} \in \mathcal{J}(\Phi)^n$ and all formulas $\varphi \in \Phi$, it is the case that $N_\varphi^{\boldsymbol{J}} \subseteq N_\varphi^{\boldsymbol{J'}}$ implies $\varphi \in F(\boldsymbol{J}) \Rightarrow \varphi \in F(\boldsymbol{J'})$.*

PROOF. ($\Leftarrow$) Suppose the stated condition holds. Independence follows by observing that $N_\varphi^{\boldsymbol{J}} = N_\varphi^{\boldsymbol{J'}}$ implies both $N_\varphi^{\boldsymbol{J}} \subseteq N_\varphi^{\boldsymbol{J'}}$ and $N_\varphi^{\boldsymbol{J'}} \subseteq N_\varphi^{\boldsymbol{J}}$. Monotonicity follows when we consider two profiles that are $\{i\}$-variants of each other.

($\Rightarrow$) Suppose F is both independent and monotonic. Consider a scenario with $N_\varphi^{\boldsymbol{J}} \subseteq N_\varphi^{\boldsymbol{J'}}$ and $\varphi \in F(\boldsymbol{J})$. We need to show $\varphi \in F(\boldsymbol{J'})$. In case $N_\varphi^{\boldsymbol{J}} = N_\varphi^{\boldsymbol{J'}}$, the claim follows from independence. Otherwise, let $\{i_1, \ldots, i_\ell\} = N_\varphi^{\boldsymbol{J'}} \setminus N_\varphi^{\boldsymbol{J}}$, with $\ell > 0$, be the agents who change their mind on φ as we move from $\boldsymbol{J}$ to $\boldsymbol{J'}$. Define a sequence of complete and consistent profiles as follows:

[1] While much of the theoretical literature on JA is concerned with the challenge of preserving complex logical dependencies during aggregation, many practical applications of JA can be accurately modelled using atomic agendas. An example is collective annotation using crowdsourcing [19].

[2] We will make use of such restrictions in two cases. In one case $\Psi^+ = \Phi^+$, and in the other Φ is assumed to be atomic. Thus, restrictions are well-defined in both cases. In the general case, care needs to be taken as not all consistent judgment sets over Ψ will necessarily remain consistent when extended with the fixed choices made for formulas in $\Phi \setminus \Psi$.

$$\begin{aligned} \boldsymbol{J}^0 &:= \boldsymbol{J} \\ \boldsymbol{J}^k &:= (\boldsymbol{J}^{k-1}_{-i_k}, J'_{i_k}) \quad \text{for } k \in \{1,\dots,\ell\} \end{aligned}$$

Thus, we start with $\boldsymbol{J}$ and replace the judgment sets of one agent at a time to eventually arrive at $\boldsymbol{J}^\ell = \boldsymbol{J'}$. By monotonicity, $\varphi \in F(\boldsymbol{J}^{k-1}) \Rightarrow \varphi \in F(\boldsymbol{J}^k)$ for all $k \in \{1,\dots,\ell\}$. Hence, by induction, we get $\varphi \in F(\boldsymbol{J}^\ell) = F(\boldsymbol{J'})$. □

Third, an aggregation rule F is *neutral* if $N^{\boldsymbol{J}}_\varphi = N^{\boldsymbol{J}}_\psi$ implies $\varphi \in F(\boldsymbol{J}) \Leftrightarrow \psi \in F(\boldsymbol{J})$. Thus, a neutral rule treats all propositions symmetrically, in the sense of either accepting all or none of any set of formulas with exactly the same coalition of supporters.

This standard axiom of neutrality, however, does not capture all symmetry requirements that one might reasonably want to impose. Specifically, as two complementary formulas φ and $\neg\varphi$ can never have the same coalition of supporters in a given profile (as accepting both would be inconsistent and as accepting neither would be incomplete), neutrality does not have to say anything about their relative treatment by the aggregation rule. To address this issue, we introduce a new axiom. We call an aggregation rule F *unbiased* if $F(\boldsymbol{J}^{\rightleftarrows S}) = F(\boldsymbol{J})^{\rightleftarrows S}$ for any profile $\boldsymbol{J} \in \mathcal{J}(\Phi)^n$ and set $S \subseteq \Phi^+$ such that also $\boldsymbol{J}^{\rightleftarrows S} \in \mathcal{J}(\Phi)^n$. Thus, unbiasedness of F requires that when we flip the formulas in the input, then the formulas in the output returned by F should flip in the same way. In other words, F is not biased for or against negated formulas.[3]

The following technical lemma will be useful later on.

Lemma 2. *Let F be an unbiased aggregation rule. Then $|F(\boldsymbol{J})| = |F(\boldsymbol{J}^{\rightleftarrows S})|$ for any profile $\boldsymbol{J} \in \mathcal{J}(\Phi)^n$ and any set $S \subseteq \Phi^+$ such that $\boldsymbol{J}^{\rightleftarrows S} \in \mathcal{J}(\Phi)^n$.*

Proof. We get $|F(\boldsymbol{J})| = |F(\boldsymbol{J})^{\rightleftarrows S}|$ as $F(\boldsymbol{J})$ is a judgment set, and $|F(\boldsymbol{J})^{\rightleftarrows S}| = |F(\boldsymbol{J}^{\rightleftarrows S})|$ as F is unbiased. □

As is well known, all uniform quota rules are independent, monotonic, and neutral [4].[4] It is easy to check that they also are unbiased, and that the same holds for the dictatorships. Constant rules are independent and monotonic, but (with the exception of the trivial rules that accept all formulas or reject all formulas) they are neither neutral nor unbiased.

3. SINGLE-AGENT MANIPULATION

In this section, we establish a baseline result on strategic manipulation by a single agent and discuss the relationship between our result and a closely related result by Dietrich and List [5]. We begin by fixing a notion of preference over judgment sets, which we will also make use of later on when discussing strategic manipulation by groups of agents.

[3]Our unbiasedness axiom is related to a different axiom by the same name used by Grossi and Pigozzi [13], which itself is a variant of the *acceptance-rejection neutrality* axiom introduced by Dietrich and List [6]. Both of these axioms are considerably stronger than unbiasedness as defined here, because we require the judgments on all non-flipped formulas to be identical in the two profiles, which these authors do not, meaning that they obtain axioms that in spirit are much closer to independence. On the other hand, we allow for more than one formula to get flipped, which they do not, so in that sense our axiom is moderately stronger.

[4]The same is not true for many other practically useful aggregation rules. In particular, many of them, such as the distance-based rules [18], violate independence.

3.1 Preferences

To model the incentives of an agent to engage in strategic manipulation, we need to model her preferences. Following Dietrich and List [5], we will assume that the most preferred judgment set of agent i is always her own truthfully held judgment set J_i and that other judgment sets are ranked in terms of their distance to J_i. The *Hamming distance* between two judgment sets $J, J' \in 2^\Phi$ is defined as the number of formulas on which they disagree:

$$H(J, J') = |J \setminus J'| + |J' \setminus J|$$

For a given profile $\boldsymbol{J} = (J_1,\dots,J_n)$, each J_i thus induces a weak order $\succcurlyeq^{\boldsymbol{J}}_i$ on judgment sets:

$$J \succcurlyeq^{\boldsymbol{J}}_i J' \quad \Leftrightarrow \quad H(J, J_i) \leqslant H(J', J_i)$$

The strict part $\succ^{\boldsymbol{J}}_i$ of $\succcurlyeq^{\boldsymbol{J}}_i$ is defined in the usual manner, i.e., $J \succ^{\boldsymbol{J}}_i J'$ if and only if $J \succcurlyeq^{\boldsymbol{J}}_i J'$ but not $J' \succcurlyeq^{\boldsymbol{J}}_i J$. All results in this paper apply to this specific type of modelling preferences, sometimes called *Hamming-distance preferences*, even when we do not always state this explicitly.

At the time of writing, this is the only standard model of preferences in the JA literature. On the other hand, it is widely recognised that this model has its limitations and that identifying richer models of preferences constitutes an important future research direction for the field. Hamming-distance preferences belong to the family of *closeness-respecting preferences*, also introduced by Dietrich and List [5], which merely demand that, for two judgment sets in $\mathcal{J}(\Phi)$, J should be (weakly) preferred to J' by agent i if $J_i \cap J \supseteq J_i \cap J'$. Some of the complexity results of Baumeister et al. [2] apply to all preference models in this class.

3.2 Strategyproofness

If agent i strictly prefers the outcome returned by F when she reports J'_i rather than her true judgment set J_i, then she has an incentive to do so, and we say that F is *susceptible* to strategic manipulation. Otherwise, we say that F is *immune* to strategic manipulation, or simply *strategyproof*.

Definition 1. *A rule F is called **strategyproof**, if for all profiles $\boldsymbol{J} \in \mathcal{J}(\Phi)^n$, agents $i \in \mathcal{N}$, and judgment sets $J'_i \in \mathcal{J}(\Phi)$ it is the case that $F(\boldsymbol{J}) \succcurlyeq^{\boldsymbol{J}}_i F(\boldsymbol{J}_{-i}, J'_i)$.*

We will sometimes write 'single-agent strategyproof' rather than just 'strategyproof' to differentiate the property from group-strategyproofness discussed later on. Every aggregation rule F that is independent and monotonic is also single-agent strategyproof. This follows from a result due to Dietrich and List [5], and may also be understood independently of their result: By independence, we can focus on one formula φ at a time. By monotonicity, if an agent wants φ to get accepted, it is in her best interest to accept φ herself (and accordingly, if she wants φ to get rejected).

The following is an immediate corollary to this result, when taken together with the observation that the uniform quota rules are both independent and monotonic, which is also due to Dietrich and List [4].

Proposition 3 (Dietrich & List, 2007). *Every uniform quota rule F_q is single-agent strategyproof.*

Dietrich and List are sometimes misquoted as also having proved that strategyproofness implies independence and monotonicity. In fact, this is not the case, as the following counterexample demonstrates.

EXAMPLE 1. *Consider an aggregation problem with pre-agenda $\Phi^+ = \{p, q\}$ and a rule F for which the outcome only depends on agent 1. F returns $\{p, q\}$ in case agent 1 accepts p, and F returns $\{\neg p, \neg q\}$ in case agent 1 accepts $\neg p$. Thus, F does not take agent 1's judgment on q into account. F is neither independent (because the acceptance of q depends on the acceptance of p) nor monotonic (because agent 1 can switch from rejecting to accepting q, with the result being that the collective switches from accepting to rejecting q). Nevertheless, F is strategyproof: in all four possible situations, agent 1 cannot increase the number of agreements between the outcome and her sincere judgment set by switching to a different judgment set.*

Hence, the set of independent and monotonic rules is a strict subset of the set of single-agent strategyproof rules. What Dietrich and List [5] *did* prove, however, is that an aggregation rule F is independent and monotonic if and only if F is strategyproof for agents with *any* kind of closeness-respecting preferences—rather than just for those with Hamming-distance preferences. For agents with Hamming-distance preferences, so far no characterisation of the class of strategyproof aggregation rules has been known. We provide such a result here that characterises the strategyproof rules *within* the class of neutral-unbiased rules.

THEOREM 4. *A neutral and unbiased aggregation rule F is single-agent strategyproof if and only if it is both independent and monotonic.*

PROOF. As we have seen earlier, the right-to-left direction holds even without the assumptions of neutrality and unbiasedness. It is a direct consequence of the aforementioned result by Dietrich and List [5]. For the other direction, we shall establish the contrapositive. So let F be neutral and unbiased, but not both independent and monotonic. We need to find a profile where F can be manipulated.

By Lemma 1, there is a situation where $N_\varphi^{\boldsymbol{J}} \subseteq N_\varphi^{\boldsymbol{J'}}$ and $\varphi \in F(\boldsymbol{J})$, but $\varphi \notin F(\boldsymbol{J'})$. When moving from $\boldsymbol{J}$ to $\boldsymbol{J'}$, one agent must be the first to trigger this change, so w.l.o.g., we may assume only agent i changed her judgment set between the two, i.e., $\boldsymbol{J}$ and $\boldsymbol{J'}$ are $\{i\}$-variants.

First, consider the special case where i is the only agent in the group (i.e., where $n = 1$). Let $S \subseteq \Phi^+$ be the set of pre-agenda formulas on which J_i and J_i' differ, i.e., $J_i^{\rightleftarrows S} = J_i'$ and thus $\boldsymbol{J}^{\rightleftarrows S} = \boldsymbol{J'}$. Hence, by unbiasedness, $F(\boldsymbol{J'}) = F(\boldsymbol{J})^{\rightleftarrows S}$. As $F(\boldsymbol{J})$ and $F(\boldsymbol{J'})$ differ on φ, this means that S must contain φ (or φ', in case φ is of the form $\varphi = \neg\varphi'$). Thus, also J_i and J_i' differ on φ. In other words, this excludes the case of $N_\varphi^{\boldsymbol{J}} = N_\varphi^{\boldsymbol{J'}}$. So we must have $N_\varphi^{\boldsymbol{J}} \subset N_\varphi^{\boldsymbol{J'}}$, meaning that $\varphi \notin J_i$ and $\varphi \in J_i'$. As $n = 1$, for any formula $\psi \in \Phi$, the coalitions of supporters of ψ in the two profiles, $N_\psi^{\boldsymbol{J}}$ and $N_\psi^{\boldsymbol{J'}}$, are either empty or the singleton $\{i\}$. This allows us to apply neutrality in the following way: If agent i agrees on φ and ψ, then also the outcome must agree on φ and ψ. Now partition Φ into four sets: the formulas agent i accepts in both profiles ($J_i \cap J_i'$), the formulas she rejects in both profiles ($\Phi \setminus (J_i \cup J_i')$), the formulas she initially rejects and then accepts ($J_i' \setminus J_i$), and those she initially accepts and then rejects ($J_i \setminus J_i'$). This perspective allows us to fully determine which formulas are either accepted or rejected in each of the two profiles:

	$F(\boldsymbol{J})$	$F(\boldsymbol{J'})$
$J_i \cap J_i'$	**out** (by Lemma 2)	**out** (by neutrality)
$\Phi \setminus (J_i \cup J_i')$	**in** (by neutrality)	**in** (by Lemma 2)
$J_i' \setminus J_i$	**in** (by neutrality)	**out** (by neutrality)
$J_i \setminus J_i'$	**out** (by Lemma 2)	**in** (by Lemma 2)

All table entries that are marked 'by neutrality' follow from neutrality together with what we know about the acceptance/rejection of φ. The remaining four table entries then follow from Lemma 2, according to which $|F(\boldsymbol{J})| = |F(\boldsymbol{J'})|$, because no other option would permit us to preserve the cardinality of the outcome when moving between the two profiles. To see this, observe (*i*) that $J_i, J_i' \in \mathcal{J}(\Phi)$ entails $|J_i' \setminus J_i| = |J_i \setminus J_i'|$ and that (*ii*) all formulas coverd by any given table entry must all behave in the same way (again, due to neutrality). In conclusion, we obtain $F(\boldsymbol{J}) = [\Phi \setminus (J_i \cup J_i')] \cup [J_i' \setminus J_i] = \Phi \setminus J_i$ and $F(\boldsymbol{J'}) = \Phi \setminus J_i'$. In other words, agent i has a clear incentive to manipulate when in profile $\boldsymbol{J}$ and to report J_i' instead of J_i.

Now suppose $n > 1$. Instead of showing directly that F can be manipulated, we instead show that if F were to be strategyproof, then there would exist another rule F' for single-agent profiles that is neutral, unbiased, not both independent and monotonic, and yet strategyproof. As we have just seen that the latter is not the case, the former cannot be the case either. So suppose F is strategyproof. We construct $F' : \mathcal{J}(\Phi)^1 \to 2^\Phi$ as follows. For any $J \in \mathcal{J}(\Phi)^1$, let $F' : (J) \mapsto F(\boldsymbol{J}_{-i}, J)$. That is, F' is the restriction of F to agent i, with the judgments of the other agents being fixed as in $\boldsymbol{J}$ and $\boldsymbol{J'}$ (recall that $\boldsymbol{J}_{-i} = \boldsymbol{J'}_{-i}$). Neutrality, unbiasedness, and strategyproofness immediately transfer from F to F', for any restriction of this kind. Only for the property of *not* being both independent and monotonic, we need to verify that the characterising property of Lemma 1 really is violated for F' as well. But this is easily seen to be the case: we have $N_\varphi^{(J_i)} \subseteq N_\varphi^{(J_i')}$ and $\varphi \in F'((J_i)) = F(\boldsymbol{J})$ but $\varphi \notin F'((J_i')) = F(\boldsymbol{J'})$. This concludes the proof. □

As an aside, we note that Theorem 4 is not in conflict with the classical Gibbard-Satterthwaite Theorem [12, 21] in voting theory, which—loosely speaking—says that no non-dictatorial voting rule for three or more alternatives can be strategyproof. The reason is that here we are lacking a counterpart for those three alternatives. Rather, in the context of the independence axiom, JA may be seen as a series of elections with two (i.e., fewer than three) alternatives, one for each formula φ in the pre-agenda, with the two alternatives being φ and $\neg\varphi$. To obtain results in JA that resemble the Gibbard-Satterthwaite Theorem one has to focus on aggregation problems with specific agendas to model preferences. Dietrich and List [5] establish a result of this kind.

4. GROUP MANIPULATION

In this section, we generalise the definition of strategyproofness discussed earlier to obtain a definition of group-strategyproofness, requiring that no coalition of agents should ever have an incentive to misreport their judgments. We then characterise the family of (neutral and unbiased) aggregation rules that is group-strategyproof in this sense.

4.1 Definition of the Concept

As we will see, the *size* of the coalition of manipulators is an important parameter in our characterisation of group-

strategyproof rules. For instance, a given rule may be subject to manipulation by three agents, but not by one or two agents. We therefore formulate our definition relative to a number k, the maximum number of agents that may form a coalition for the purposes of strategic manipulation.

DEFINITION 2. *An aggregation rule F is called **group-strategyproof** against coalitions of up to k manipulators, if for all profiles $\boldsymbol{J} \in \mathcal{J}(\Phi)^n$, coalitions $C \subseteq \mathcal{N}$ with $|C| \leqslant k$, and C-variants $\boldsymbol{J'} \in \mathcal{J}(\Phi)^n$ of $\boldsymbol{J}$ it is the case that $F(\boldsymbol{J}) \succcurlyeq_i^{\boldsymbol{J}} F(\boldsymbol{J'})$ for all agents $i \in C$.*

In the above definition, C is the set of manipulators and the other agents are truthful. $\boldsymbol{J}$ is the truthful profile and $\boldsymbol{J'}$ is the result of the agents in C misreporting their judgments. Intuitively, the definition says that there exists no group of up to k agents who can, in at least one situation, change their judgment sets in such a way that they all (strictly) prefer the new outcome over the old one. Note that for the special case of $k = 1$, we recover Definition 1.

EXAMPLE 2. *Consider the following profile with five agents and the pre-agenda $\Phi^+ = \{\varphi_1, \varphi_2, \varphi_3\}$:*

	φ_1	φ_2	φ_3	$\neg\varphi_1$	$\neg\varphi_2$	$\neg\varphi_3$
Agent 1	×	✓	✓	✓	×	×
Agent 2	✓	×	✓	×	✓	×
Agent 3	✓	✓	×	×	×	✓
Agent 4	×	×	×	✓	✓	✓
Agent 5	×	×	×	✓	✓	✓
Majority	×	×	×	✓	✓	✓

Now suppose the first three agents manipulate by flipping their judgments corresponding to the cells shaded in grey. This will cause every single formula in the majority outcome to flip as well. In the initial profile, for each of the manipulators, the Hamming distance between the outcome and her individual judgment set is 4. After the change, the distance between the new outcome and the old (i.e., truthful) individual judgment set shrinks to 2 for each of them. Hence, the majority rule is not group-strategyproof against three manipulators—at least not in a world with a population of five. Importantly, observe that each manipulator performs a manipulation on one formula (and its complement) that damages her own interests, but this is made up for by the other two performing manipulations that benefit her.

Two families of aggregation rules that are immediately seen to be group-strategyproof, against coalitions of any size, are the dictatorships and the constant rules.

Observe that group-strategyproofness for (up to) k implies group-strategyproofness for (up to) $k-1$. Thus, for positive results we will look for the largest k for which they apply, while for negative results we will look for the smallest such k. If F is group-strategyproof for any coalition size k, we simply say that F is group-strategyproof.

4.2 Manipulation by Two Agents

We first analyse the case of $k = 2$, i.e., the case of up to two agents manipulating, and find that the situation is no worse than for the single-agent case covered by Theorem 4.

THEOREM 5. *A neutral and unbiased aggregation rule F is group-strategyproof against coalitions of up to two manipulators if and only if it is both independent and monotonic.*

PROOF. ($\Rightarrow$) We proceed by contraposition. Suppose F is neutral and unbiased but not both independent and monotonic. Then, by Theorem 4, F is not group-strategyproof even against a single agent (i.e., a coalition of one manipulator). Hence, F certainly is not group-strategyproof for coalitions of *up to* two manipulators.

($\Leftarrow$) Assume F is neutral, unbiased, independent, and monotonic. By Theorem 4, there cannot be a manipulation with a coalition of a single agent. So, w.l.o.g., suppose there are *exactly* two manipulators. Since F is monotonic, the manipulators cannot manipulate on formulas for which their individual judgment sets agree (because their truthful judgments on such formulas are optimal for both of them already). Therefore, any manipulator who changes her individual judgment set to have an effect on the outcome must go against her own preferences in order to benefit the other agent. As preferences are defined in terms of the Hamming distance, each of the two manipulators will have to require the other one to perform *strictly more* such manipulations on individual formulas than she performs herself. But this clearly is a contradiction. □

4.3 Manipulation by More Than Two Agents

To understand the case of arbitrary numbers of manipulators, we first focus on the case of $k = 3$ manipulators. We will require a definition of what it means for an agent to be able to have an effect on a given formula in a given profile.

DEFINITION 3. *Agent $i \in \mathcal{N}$ is called **effective** on formula $\varphi \in \Phi^+$ in profile $\boldsymbol{J} \in \mathcal{J}(\Phi)^n$ with $J_i^{\rightleftarrows\varphi} \in \mathcal{J}(\Phi)$ if it is the case that both $\varphi \in F(\boldsymbol{J}) \not\Leftrightarrow \varphi \in F(\boldsymbol{J}_{-i}, J_i^{\rightleftarrows\varphi})$ and $\neg\varphi \in F(\boldsymbol{J}) \not\Leftrightarrow \neg\varphi \in F(\boldsymbol{J}_{-i}, J_i^{\rightleftarrows\varphi})$.*

Thus, agent i is effective on φ in $\boldsymbol{J}$, if she can flip her judgment on φ whilst remaining complete and consistent, and if due to that flip the status of both φ and $\neg\varphi$ changes in the outcome. If, for the initial outcome $F(\boldsymbol{J})$, it is the case that exactly one of φ and $\neg\varphi$ is accepted (i.e., if F is complete and complement-free in this instance), then this property is preserved in the changed outcome $F(\boldsymbol{J}_{-i}, J_i^{\rightleftarrows\varphi})$. Otherwise, it continues to be violated.

We now prove a technical lemma that fully characterises the class of group-strategyproof aggregation rules for the special case of three agents (and thus at most three manipulators) and an agenda of three positive formulas (i.e., six formulas overall). We will later use this characterisation as a gadget to characterise rules for arbitrary numbers of agents and arbitrary agendas that are group-strategyproof against up to three manipulators. The first part of the proof of the following lemma amounts to a generalisation of the technique employed in Example 2 to obtain a profile in which a successful manipulation may occur.

LEMMA 6. *Suppose $\mathcal{N} = \{1, 2, 3\}$ and $|\Phi^+| = 3$. Then a neutral and unbiased aggregation rule F is group-strategyproof if and only if it is independent and monotonic and there are no unanimous profile $(J, J, J) \in \mathcal{J}(\Phi)^3$ and bijection $g : \mathcal{N} \to \Phi^+$ such that, for all $i \in \mathcal{N}$, agent i is effective on agenda formula $g(i)$ in profile $(J^{\rightleftarrows g(1)}, J^{\rightleftarrows g(2)}, J^{\rightleftarrows g(3)}) \in \mathcal{J}(\Phi)^3$.*

PROOF. ($\Rightarrow$) We prove the contrapositive. Let F be a neutral and unbiased aggregation rule. First of all, if F is not both independent and monotonic, we know that it is

manipulable by a single agent by Theorem 4. Therefore, assume F is independent and monotonic, and suppose that there are a unanimous profile (J, J, J) and a bijection g as described above. Let $\Phi^+ = \{\varphi_1, \varphi_2, \varphi_3\}$. W.l.o.g., as we can rename the formulas, we may assume that $g(i) = \varphi_i$ for all $i \in \mathcal{N}$. Also w.l.o.g., we may assume that $J = \{\varphi_1, \varphi_2, \varphi_3\}$, because for any aggregation problem where a negative formula $\neg\varphi_i$ occurs in J, there exists an equivalent aggregation problem where $\neg\varphi_i$ has been replaced by the positive formula $\top \wedge \neg\varphi_i$ and φ_i by its negation. We now need to show that there exists a profile in which manipulation can occur. This profile is $\boldsymbol{J} = (J^{\rightleftarrows\varphi_1}, J^{\rightleftarrows\varphi_2}, J^{\rightleftarrows\varphi_3})$. Indeed, let $\boldsymbol{J}$ be the truthful profile. Then the result of each agent i manipulating by flipping formula φ_i is profile (J, J, J). As all three agents are effective, all three formulas must have changed status during manipulation. By monotonicity, this change must have occurred "in the right direction", i.e., we have $F(\boldsymbol{J}) = \{\neg\varphi_1, \neg\varphi_2, \neg\varphi_3\}$ and $F(J, J, J) = \{\varphi_1, \varphi_2, \varphi_3\}$.

We can now easily verify that for each agent i the Hamming distance between her truthful judgment set $J^{\rightleftarrows\varphi_i}$, e.g., $\{\neg\varphi_1, \varphi_2, \varphi_3\}$ for agent 1, and the nonmanipulated outcome $\{\neg\varphi_1, \neg\varphi_2, \neg\varphi_3\}$ is higher than the Hamming distance between her truthful judgment set and the manipulated outcome $\{\varphi_1, \varphi_2, \varphi_3\}$, namely $H(J^{\rightleftarrows\varphi_i}, F(\boldsymbol{J})) = 4$ rather than $H(J^{\rightleftarrows\varphi_i}, F(J, J, J)) = 2$. Therefore, F is not group-strategyproof against manipulation by all three agents.

($\Leftarrow$) Let F be an aggregation rule that is neutral, unbiased, independent, and monotonic. Now suppose that for every unanimous profile (J, J, J) and for every bijection g there is some agent i such that i is not effective on formula $g(i)$. This amounts to saying that, for every unanimous profile (J, J, J) and every possible bijection g, there will always be at most two effective agents over the formulas in Φ^+. Hence, it will always be possible to use the reasoning employed in the proof of the right-to-left direction of Theorem 5 to show that F must be group-strategyproof. □

Observe that independence and neutrality of an aggregation rule F together imply that F must be induced by a local rule, a boolean function over subsets of $\mathcal{N}$, that decides for each formula φ in the agenda whether it should be accepted based only on the coalitions of individual agents that accept φ. This observation is useful in that it allows us to fully classify all aggregation rules covered by Lemma 6 as either group-strategyproof or susceptible to strategic manipulation. For three agents and one formula (that may be accepted or rejected), there are $2^{2\cdot2\cdot2} = 256$ possible local rules. But only 20 of them are monotonic.[5] We have enumerated these rules using a simple computer program. They consist of 3 *dictatorships* (one for each agent), 2 *constant rules* (always-accept and always-reject), the *nomination rule* (the quota rule with quota 1), the *majority rule* (the quota rule with quota 2), the *unanimity rule* (the quota rule with quota 3), 6 *two-agent rules* (with one dummy agent and the other two using either a nomination or a unanimity rule), and 6 *weighted quota rules* (with one agent having weight 2, two having weight 1, and the quota being either 2 or 3).

Using Lemma 6 it is now easy to check that, for $n = 3$ and $|\Phi^+| = 3$, amongst the neutral, unbiased, independent, and monotonic aggregation rules, the only ones that can be manipulated by the full group are the nomination rule and the unanimity rule. For example, for the nomination rule all three agents are effective for φ in any situation where currently none of them accepts φ. On the other hand, for the majority rule, for instance, it is impossible to find a situation where all three agents are effective for φ whilst currently agreeing on φ (agent 1 is effective only when agents 2 and 3 disagree, and so forth). Thus, we may rewrite Lemma 6 in the following simplified form.

LEMMA 7. *Suppose $n = 3$ and $|\Phi^+| = 3$. Then a neutral and unbiased aggregation rule F is group-strategyproof if and only if it is independent and monotonic, and if it is neither the nomination rule nor the unanimity rule.*

Unfortunately, as Example 2 demonstrates, this very positive result does not generalise to larger numbers of agents. Specifically, although the majority rule is group-strategyproof for three agents, it can be manipulated by three agents when the overall number of agents is five. The reason is that, when we use the majority rule for five agents and keep the judgments of two (sincere) agents fixed, then the resulting rule implicitly defined over the remaining three agents is in fact the nomination rule, which we have seen to be manipulable in a world with only three agents.

For our next result we are going to generalise this idea of a rule for three agents being implicitly defined by a rule for a larger population when we keep the judgments of all but three agents fixed. Note that this result only applies to the case of atomic agendas; the implications of this restriction are discussed at the end of this section.

THEOREM 8. *Suppose the agenda Φ is atomic. Then a neutral and unbiased aggregation rule F is group-strategyproof against coalitions of up to three manipulators if and only if F is independent and monotonic, and if none of the restrictions of F to three agents and three pre-agenda formulas is either the nomination rule or the unanimity rule.*

PROOF. By Theorem 4, amongst the neutral and unbiased aggregation rules, only the independent and monotonic rules can potentially be group-strategyproof. Thus, what the theorem claims on top of this basic insight is that a neutral, unbiased, independent, and monotonic rule F can be manipulated by a coalition of 3 agents *if and only if* there exist a set of 3 agents and a set of 3 formulas in the pre-agenda such that, when we keep everything else fixed (the judgments of the other $n - 3$ agents on everything and the judgments of the 3 selected agents on the other agenda formulas), the resulting aggregation rule is either the nomination rule or the unanimity rule. The right-to-left direction of this latter claims is an immediate consequence of Lemma 7: If there is such a restriction that is the nomination or the unanimity rule, then that restricted rule can be manipulated by Lemma 7, and thus F can be manipulated by the very same moves. The left-to-right direction of the same claim is also immediate in case the original aggregation problem has a pre-agenda of 3 formulas only.

So all that remains to be shown is that the focus on 3 pre-agenda formulas at a time is sufficient to catch at least one situation in which manipulation is possible whenever manipulation is possible at all. So suppose that in a world with 3 agents there exists a successful manipulation involving m positive formulas. We need to show that then there also

[5]Readers familiar with coalitional game theory may recognise 20, the third Dedekind number [11], as the number of simple games for three players [22].

exists a manipulation involving only 3 of those formulas.[6] First, w.l.o.g., we may assume that every formula involved in the manipulation changed its acceptance status in the outcome (otherwise, simply not manipulating such a formula does not change the result). Second, w.l.o.g., we may assume that none of the formulas involved in the manipulation was flipped by all 3 agents (because otherwise none of them would have benefitted from getting this formula flipped in the outcome). So every formula involved is flipped by either one or two agents. The agents flipping φ always prefer the sincere outcome for φ, and the ones not flipping φ always prefer the insincere outcome for φ (just as in Example 2). Thus, there are six types of formulas: $(+1,+1,-1)$-formulas are those where the manipulation benefits agents 1 and 2 but harms agent 3; $(+1,-1,-1)$-formulas are those where the manipulation benefits 1 but harms 2 and 3; and so forth. The numbers, $+1$ and -1, indicate the changes to the Hamming distances for the 3 agents caused by the formula in question. When adding up the vectors corresponding to the m formulas we must get a positive number in each of the 3 positions (otherwise the corresponding agent does not benefit from the overall manipulation). It is easy to see that this is only possible if of $(+1,+1,-1)$, $(+1,-1,+1)$, and $(-1,+1,+1)$, i.e., the formulas benefitting two agents and harming only one agent, each shows up at least once. But if the agents only manipulate on the corresponding 3 formulas, they also all benefit, i.e., we have shown that a manipulation involving only 3 positive formulas is feasible. □

We must emphasise that Theorem 8 is a negative result, as it implies that essentially all reasonable (neutral and unbiased) aggregation rules are susceptible to strategic manipulation by groups of agents. To substantiate this claim, let us see what Theorem 8 implies for the large and important family of the the uniform quota rules.

COROLLARY 9. *No uniform quota rule F_q with a quota q satisfying $3 \leqslant q \leqslant n$ or $1 \leqslant q \leqslant n-2$ that is defined on an atomic agenda Φ is group-strategyproof.*

PROOF. All uniform quota rules are neutral, unbiased, independent, and monotonic. Thus, by Theorem 8, to show a failure of group-strategproofness, we must find a restriction of F_q to 3 agents and 3 pre-agenda formulas for which the local rule deciding on the acceptance of each individual formula is either the nomination or the unanimity rule.

First, suppose $3 \leqslant q \leqslant n$. In this case, we can find a restriction of F_q to 3 agents and 3 pre-agenda formulas that is the unanimity rule: simply consider a profile in which $q-3$ sincere agents accept φ and $n-q$ sincere agents reject φ. Then φ will get accepted if and only if the three manipulators unanimously accept φ. Note that, due to our assumptions on q, this construction is well-defined: $q-3 \geqslant 0$, $n-q \geqslant 0$, and $(q-3)+(n-q)+3 = n$.

Second, suppose $1 \leqslant q \leqslant n-2$. In this case, we can find a restriction of F_q that is the nomination rule in an analogous manner: consider a profile in which $q-1$ sincere agents accept φ and $n-2-q$ sincere agents reject φ. Then φ will get accepted if and only if at least one of the manipulators accepts it. Also this construction is well-defined: $q-1 \geqslant 0$, $n-2-q \geqslant 0$, and $(q-1)+(n-2-q)+3 = n$. □

[6]Because of our assumption that Φ is atomic, we are free to change the judgments for the remaining $m-3$ formulas back to what they were in the sincere judgment sets. This is the only reason for this assumption.

The assumptions on the quota under which Corollary 9 applies are very weak. Observe that $n > 3$ implies that either $3 \leqslant q$ or $q \leqslant n-2$ (or both). Thus, for aggregation problems with $n > 3$ agents, the only uniform quota rules that are group-strategyproof are the trivial rules that either accept or reject all formulas. These are the rules we obtain for $q = 0$ and $q = n+1$, respectively. For the special case of $n = 3$ agents, as we have seen in the discussion following Lemma 6, in addition the majority rule is group-strategyproof. For $n < 3$ agents, all uniform quota rules are group-strategyproof. This follows from Theorem 5.

Theorem 8 is stated for the case of (up to) $k = 3$ manipulators, rather than for arbitrary k. As pointed out in Section 4.1, as it is essentially a negative result (i.e., the contrapositive of the left-to-right direction, giving sufficient conditions for manipulability, is what matters most), this is the strongest possible form of this kind of result. Indeed, as essentially every reasonable neutral rule may be manipulated by coalitions of up to 3 manipulators, it certainly is the case that they also are subject to manipulation by coalitions of up to $k = 4, 5, \ldots$ manipulators.

Finally, Theorem 8 only applies to atomic agendas. We required this restriction to be able to show that any manipulation involving m positive formulas can be reduced to a manipulation involving 3 positive formulas. For an agenda with strong logical dependencies between formulas, more positive results are possible in principle. For example, in the most extreme case, when an agenda consists of tautologies and their negations only (and assuming *unanimity*),[7] strategic manipulation is impossible, because every agent will have the exact same sincere judgment set.[8] However, for any agenda permitting a fair degree of variation in the range of admissible judgment sets, our results suggest that group-strategyproofness is rarely an option. A full analysis of how agenda conditions impact on group-strategyproofness constitutes an important direction for future work.

5. UNSTABLE MANIPULATION

In this section, we propose a variant of the notion of group-strategyproofness defined earlier and show that under this new definition much more positive results are feasible. This alternative definition is inspired by the following example.

EXAMPLE 3. *Recall the scenario discussed in Example 2, where five agents were using the majority rule to decide on three pairs of formulas, and the first three agents had an incentive to manipulate. Below is the profile $\boldsymbol{J'}$ we obtain after they manipulate as indicated before (starting from $\boldsymbol{J}$):*

	φ_1	φ_2	φ_3	$\neg\varphi_1$	$\neg\varphi_2$	$\neg\varphi_3$
Agent 1	✓	✓	✓	×	×	×
Agent 2	✓	✓	✓	×	×	×
Agent 3	✓	✓	✓	×	×	×
Agent 4	×	×	×	✓	✓	✓
Agent 5	×	×	×	✓	✓	✓
Majority	✓	✓	✓	×	×	×

Consider what happens when agent 1 opts out of the joint plan and reverses her manipulation, i.e., when we flip the

[7]The unanimity axiom states that any reasonable aggregation rule should map unanimous profiles $(J, \ldots, J)$ to J.

[8]Recall that individual judgment sets are always complete and consistent, i.e., all agents accept all tautologies.

judgments corresponding to the cells shaded in grey. Then the majority judgment on φ_1 and $\neg\varphi_1$ changes as well. As a result, agent 1 will obtain her most preferred outcome $F(\boldsymbol{J'}_{-1}, J_1) = \{\neg\varphi_1, \varphi_2, \varphi_3\}$, which is equal to her truthful judgment set. So agent 1's preferences in our example are actually $F(\boldsymbol{J'}_{-1}, J_1) \succ_1^{\boldsymbol{J}} F(\boldsymbol{J'}) \succ_1^{\boldsymbol{J}} F(\boldsymbol{J})$. Thus, she has an incentive to opt out of the joint plan—and for symmetry reasons, so does every other agent.[9]

This is in contrast with our previous notion of manipulation, where any agreed-upon manipulation is guaranteed to be carried out by all agents in the coalition of manipulators. Now coalitions are *fragile* in the sense that any one agent may opt out of a previously agreed upon manipulation, if doing so is in her own interest.

DEFINITION 4. *An aggregation rule F is called* ***group-strategyproof against fragile coalitions*** *of up to k manipulators, if for all profiles $\boldsymbol{J} \in \mathcal{J}(\Phi)^n$, coalitions $C \subseteq \mathcal{N}$ with $|C| \leqslant k$, and C-variants $\boldsymbol{J'} \in \mathcal{J}(\Phi)^n$ of $\boldsymbol{J}$ with $F(\boldsymbol{J'}) \succ_i^{\boldsymbol{J}} F(\boldsymbol{J})$ and $F(\boldsymbol{J'}_{-i}, J_i) \neq F(\boldsymbol{J'})$ for all $i \in C$ it is the case that $F(\boldsymbol{J'}_{-i}, J_i) \succ_i^{\boldsymbol{J}} F(\boldsymbol{J'})$ for some $i \in C$.*

Intuitively, F is group-strategyproof against fragile coalitions if any coalition attempting manipulation is one where every agent in the coalition has an incentive to opt out of the manipulation (i.e., to go back to her truthful judgment set). In the above definition, C is the coalition of manipulators. Condition $F(\boldsymbol{J'}) \succ_i^{\boldsymbol{J}} F(\boldsymbol{J})$ says that all agents in the coalition benefit from the manipulation and condition $F(\boldsymbol{J'}_{-i}, J_i) \neq F(\boldsymbol{J'})$ says that every member of the coalition is in fact needed to achieve the new outcome. Finally, $F(\boldsymbol{J'}_{-i}, J_i) \succ_i^{\boldsymbol{J}} F(\boldsymbol{J'})$ says that agent i benefits from opting out once we are in the new profile $\boldsymbol{J'}$.

Our new definition makes manipulation more difficult and thereby immunity against manipulation more likely. Indeed, we are able to fully recover the class of aggregation rules that was shown to be single-agent strategyproof in Theorem 4, but now against coalitions of manipulators of any size, provided those coalitions are fragile.

THEOREM 10. *A neutral and unbiased aggregation rule F is group-strategyproof against fragile coalitions of manipulators if and only if it is independent and monotonic.*

PROOF. ($\Rightarrow$) Observe that F is single-agent strategyproof if and only if it is group-strategyproof against fragile coalitions of size 1 (as no single manipulator would ever want to opt out of its own coalition). Hence, group-strategyproofness against fragile coalitions of manipulators implies standard single-agent strategyproofness. The claim then follows from Theorem 4.

($\Leftarrow$) Let F be independent and monotonic. Consider two C-variants $\boldsymbol{J}, \boldsymbol{J'} \in \mathcal{J}(\Phi)^n$ such that $F(\boldsymbol{J'}) \succ_i^{\boldsymbol{J}} F(\boldsymbol{J})$ and $F(\boldsymbol{J'}_{-i}, J_i) \neq F(\boldsymbol{J'})$ for all $i \in C$. The latter means that every single agent in C is effective in their manipulation. By independence, we can reason formula-by-formula. By monotonicity, the effect that an agent i has on a formula she is effective for must be against her own true preferences. Thus, if she reverts back to her truthful judgment set, she will necessarily obtain a strictly preferred outcome. □

[9]Also consider what happens if agents 1 and 2 both opt out. Then we end up in a profile with outcome $\{\neg\varphi_1, \neg\varphi_2, \varphi_3\}$, which is the worst possible outcome for agent 3. Thus, if agent 3 is afraid that the others may opt out, she will not agree on a joint manipulation in the first place.

In other words, under the alternative notion of strategyproofness discussed in this section, and when restricting attention to neutral and unbiased aggregation rules, single-agent strategyproofness and group-strategyproofness in fact define the same class of aggregation rules.

6. CONCLUSION

We have introduced the notion of group-strategyproofness into JA and found that it is a considerably more demanding property than ordinary strategyproofness against a single manipulator. For example, we have seen that, while all uniform quota rules are single-agent strategyproof (see Proposition 3), essentially none of them are group-strategyproof (see Corollary 9). Our main result, Theorem 8, shows that, for a neutral and unbiased aggregation rule, single-agent strategyproofness extends to group-strategyproofness only in case the rule is such that it is impossible to obtain either the unanimity rule or the nomination rule by fixing the judgments of all but three agents as well as the judgments of those three agents on all but three (pairs of) formulas. That this condition is highly unlikely to be satisfied is convincingly demonstrated by the aforementioned instantiation of this result to the uniform quota rules.

We have also seen that when the contracts manipulating agents make with each other are not binding, then no coordinated group manipulation is stable. This may be interpreted as offering protection against group manipulation in practice. Other forms of protection are interesting topics for future research. For instance, as we have already indicated, strong logical dependencies in the agenda will reduce opportunities for manipulation, and narrow *domain restrictions* [6] may offer full protection. Another avenue to pursue in this context is to extend existing work on the *complexity of manipulation* [2, 8] to group manipulation.

Our characterisation result for single-agent strategyproofness is of some interest in its own right and clarifies an issue that had remained vague in the literature to date, namely that independence and monotonicity alone are not sufficient to guarantee single-agent strategyproofness when an agent's preferences are induced by the Hamming distance between the outcome and an agent's true judgment set. Our Theorem 4 shows that, within the class of the neutral and unbiased rules, single-agent strategyproofness is fully characterised by independence and monotonicity. When neutrality and unbiasedness are dropped as assumptions, it is known that independence and monotonicity still are sufficient conditions for strategyproofness, but a full characterisation of the strategyproof rules for Hamming-distance preferences remains an open problem.

We have made the common assumption that preferences are induced by the Hamming distance. Our main result is negative, so it also applies to the larger class of closeness-respecting preferences. On the other hand, our positive result on group-strategyproofness against two manipulators does not extend to closeness-respecting preferences more generally, as our proof directly exploits the properties of the Hamming distance. Nevertheless, considering other forms of preferences constitutes an important direction for future work on strategic behaviour in JA, and this includes future work on group-strategyproofness.

Acknowledgments. We would like to thank the anonymous reviewers for the helpful feedback provided.

REFERENCES

[1] K. J. Arrow, A. K. Sen, and K. Suzumura, editors. *Handbook of Social Choice and Welfare*. North-Holland, 2002.

[2] D. Baumeister, G. Erdélyi, O. J. Erdélyi, and J. Rothe. Complexity of manipulation and bribery in judgment aggregation for uniform premise-based quota rules. *Mathematical Social Sciences*, 76:19–30, 2015.

[3] F. Brandt, V. Conitzer, and U. Endriss. Computational social choice. In G. Weiss, editor, *Multiagent Systems*, pages 213–283. MIT Press, 2013.

[4] F. Dietrich and C. List. Judgment aggregation by quota rules: Majority voting generalized. *Journal of Theoretical Politics*, 19(4):391–424, 2007.

[5] F. Dietrich and C. List. Strategy-proof judgment aggregation. *Economics and Philosophy*, 23(3):269–300, 2007.

[6] F. Dietrich and C. List. Majority voting on restricted domains. *Journal of Economic Theory*, 145(2):512–543, 2010.

[7] U. Endriss. Judgment aggregation. In F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia, editors, *Handbook of Computational Social Choice*. Cambridge University Press, 2016.

[8] U. Endriss, U. Grandi, and D. Porello. Complexity of judgment aggregation. *Journal of Artificial Intelligence Research (JAIR)*, 45:481–514, 2012.

[9] P. Everaere, S. Konieczny, and P. Marquis. The strategy-proofness landscape of merging. *Journal of Artificial Intelligence Research (JAIR)*, 28(1):49–105, 2007.

[10] P. Everaere, S. Konieczny, and P. Marquis. Belief merging versus judgment aggregation. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-2015)*, 2015.

[11] R. Fidytek, A. W. Mostowski, R. Somla, and A. Szepietowski. Algorithms counting monotone Boolean functions. *Information Processing Letters*, 79(5):203–209, 2001.

[12] A. Gibbard. Manipulation of voting schemes: A general result. *Econometrica*, 41(4):587–601, 1973.

[13] D. Grossi and G. Pigozzi. *Judgment Aggregation: A Primer*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2014.

[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] L. A. Kornhauser and L. G. Sager. The one and the many: Adjudication in collegial courts. *California Law Review*, 81(1):1–59, 1993.

[16] C. List and P. Pettit. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy*, 18(1):89–110, 2002.

[17] C. List and C. Puppe. Judgment aggregation: A survey. In P. Anand, P. Pattanaik, and C. Puppe, editors, *Handbook of Rational and Social Choice*. Oxford University Press, 2009.

[18] M. K. Miller and D. Osherson. Methods for distance-based judgment aggregation. *Social Choice and Welfare*, 32(4):575–601, 2009.

[19] C. Qing, U. Endriss, R. Fernández, and J. Kruger. Empirical analysis of aggregation methods for collective annotation. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING-2014)*, 2014.

[20] I. Rahwan and F. Tohmé. Collective argument evaluation as judgement aggregation. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-2010)*, 2010.

[21] 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(2):187–217, 1975.

[22] A. D. Taylor and W. S. Zwicker. *Simple Games: Desirability Relations, Trading, Pseudoweightings*. Princeton University Press, 1999.