Empirical Evaluation of Collective Rationality for Quota Rules in Judgment Aggregation

Sharon Gieske, Elise van der Pol, and Ulle Endriss

University of Amsterdam

Abstract

A major challenge in multiagent systems research is to design good aggregation rules for combining the judgments—or beliefs, or opinions—of different autonomous software agents collaborating with each other. If the chosen rule is too sophisticated, we may encounter algorithmic difficulties. But if it is too simple, we may encounter some of the paradoxes of social choice theory and end up with inconsistent information at the system level. We report on an empirical study aimed at improving our understanding of how frequently these paradoxes strike in practice for a class of simple aggregation rules, the uniform quota rules. Our results indicate that quota rules can be expected to work significantly better in practice than the still relatively scarce theoretical results may suggest.

1 Introduction

Aggregating the opinions of a group of agents into a collective decision is a problem that arises both in the context of human decision making, e.g., in an election, and in the context of decision making in multiagent systems in artificial intelligence [13, 15]. The field of *social choice theory* [1] deals with the normative and mathematical aspects of collective decision making, and *computational social choice* [3] adds an algorithmic perspective, which is often crucial for applications in artificial intelligence. When individual opinions consist of the acceptance or rejection of several issues, and when there are some interdependencies between those issues, we speak of *judgment aggregation* [5, 8, 11]. Several rules for aggregating individual judgments have been proposed and studied in the literature. A well-known and much-used family are the (uniform) *quota rules* [4]. They accept or reject issues based on the number of supporters they have: if an issue has more supporters than a certain threshold (the quota), then it is accepted. Quota rules are attractive due to their conceptual simplicity and low computational complexity, but they also have a serious drawback: it is possible for a group of agents to submit rational judgements over a set of issues, with the aggregation rule nevertheless producing an irrational outcome.

Example 1. Suppose five intelligent robots are tasked with providing basic services in a large congress centre. They each have a sensor for measuring the temperature and they each are equipped with a camera to assess how busy the centre is. Based on this kind of information, they have to decide whether to switch on the air conditioning system. They make the following individual judgments:

	Warm	*Busy*	*A/C*
Robot 1:	Yes	Yes	Yes
Robot 2:	Yes	Yes	Yes
Robot 3:	Yes	No	No
Robot 4:	No	Yes	No
Robot 5:	Yes	No	No

How should they find a consensus? In other words, what should be the collective judgment of this multi-robot system regarding the issues *Warm*, *Busy*, and *A/C*? If they use the (weak) *majority rule*, under which an issue is accepted if at least 50% of the individuals accept it, the system will accept *Warm* and *Busy* but reject *A/C*. Now suppose we impose the integrity constraint $\mathit{Warm} \wedge \mathit{Busy} \rightarrow \mathit{A/C}$, meaning that we consider it irrational if someone accepts that it is warm and busy, but rejects the proposition that

the A/C should be switched on. Under this view, each individual robot is rational, but the system view produced by the majority rule is not. This is an instance of the famous *doctrinal paradox* [9,10]. We also say that the majority rule violates *collective rationality*, because it does not (always) preserve rationality during aggregation. Of course, the circumstances under which this paradox manifests itself are fairly narrow: For instance, if Robot 3 changes its vote to that of Robot 1, there is no problem (because *A/C* will be accepted). Alternatively, if we increase the quota required for acceptance of an issue from 50% to 67%, there also is no problem (because *Busy* will be rejected, i.e., the outcome is rational). Finally, if we change the integrity constraint to $\neg$*Warm* $\rightarrow$ $\neg$*A/C* ("if it's cold, don't switch on the A/C"), then there also is no problem (because the majority outcome satisfies this new constraint). □

How pervasive is this violation of collective rationality? Theoretical results have identified certain special cases, relating the quota to the logical structure of the integrity constraint, in which we can *guarantee* collective rationality [4, 7]. For instance, as is easy to check, in our example with integrity constraint *Warm* $\wedge$ *Busy* $\rightarrow$ *A/C*, a quota of 67% (as well as any higher quota) guarantees a rational output for *any* profile of rational inputs. Unfortunately, theoretical guarantees of this kind rely on fairly strong assumptions, and these assumptions cannot always be satisfied in real-world applications. For example, the quota required to guarantee collective rationality may be too high for the corresponding rule to be of practical interest, if such levels of consensus cannot be reached in a given group.

In this paper, we therefore ask how frequently rationality of the outcome is achieved in practice, even when it cannot be guaranteed for every single rational input profile. To this end, we have created a number of aggregation problems (represented by their integrity constraints) and generated a large number of profiles of rational judgments for a group of agents (for varying assumptions on how many random "mistakes" an agent may make on specific issues), to then determine, for a given quota q, how frequently the outcome produced by the quota rule with quota q is rational. By varying q, this approach allows us to derive recommendations for what kind of quota might be appropriate for a given aggregation problem with certain parameters, so as to ensure rational outcomes in most cases, even when a full theoretical guarantee is infeasible. To the best of our knowledge, this is the first empirical study in judgment aggregation trying to investigate the practical relevance of paradoxes, the theoretical aspects of which have been the subject of a large and still growing literature.

The remainder of this paper is organised as follows. In Section 2 we introduce the formal framework we will use. Section 3 describes how we have generated the synthetic data for our experiments, and Section 4 presents a series of such experiments and discusses our findings. Section 5 concludes.

2 Binary Aggregation with Integrity Constraints

There are a number of different formal frameworks for judgment aggregation available [5, 8, 11]. We will work with a framework known as *binary aggregation with integrity constraints* [6]. In this section, we recall the basic definitions for this framework as well as some relevant known results.

2.1 Formal Framework

Let $\mathcal{I} = \{1, \ldots, m\}$ be a finite set of binary *issues*. A *ballot* is a vector $B = (b_1, \ldots, b_m) \in \{0,1\}^m$, indicating for each issue $j \in \mathcal{I}$ whether it is *accepted* ($b_j = 1$) or *rejected* ($b_j = 0$). We associate each $j \in \mathcal{I}$ with a propositional variable p_j. To specify which ballots should be considered *rational*, we use an *integrity constraint* Γ, a formula of propositional logic over the set of variables induced by $\mathcal{I}$. For example, the integrity constraint $\Gamma = (p_1 \vee p_2)$ says that at least one of the first two issues must be accepted. We write $\mathrm{Mod}(\Gamma)$ for the set of models of Γ, i.e., $\mathrm{Mod}(\Gamma)$ is the set of rational ballots. For example, assuming $m = 3$, we get $(1,0,1) \in \mathrm{Mod}(p_1 \vee p_2)$, i.e., accepting the first and the third issue, but rejecting the second, would be rational w.r.t. $\Gamma = (p_1 \vee p_2)$.

W.l.o.g., we assume that integrity constraints are always given in conjunctive normal form (CNF), i.e., as conjunctions of clauses (possibly just a single clause). A *k-clause* is a disjunction of k literals (with no propositional variable occurring more than once). A positive k-clause (or *k-pclause* for short) is a k-clause with only positive literals; negative k-clauses (or *k-nclauses*) are defined analogously.

Let $\mathcal{N} = \{1, \ldots, n\}$ be a finite set of *agents*. A *profile* is a vector of ballots $\boldsymbol{B} = (B_1, \ldots, B_n)$, one for each agent. A (resolute) *aggregation rule* is a function $F : \mathrm{Mod}(\Gamma)^n \rightarrow \{0,1\}^m$, mapping any given profile of rational ballots into a single consensus ballot (which need not always be rational).

In this paper, we will focus on the family of *uniform quota rules*. For a given quota $q \in [0,1]$, a real number between 0 and 1, the corresponding quota rule F_q accepts issue $j \in \mathcal{I}$ if and only if at least $q \cdot n$ of the individual agents do. Thus, $F_{0.5}$ is the weak majority rule, F_1 is the unanimity rule accepting only those issues accepted by all agents, and F_0 is a trivial rule that always accepts all issues.

Example 2. We can now present our initial example more formally. There are $m = 3$ issues and $n = 5$ agents. The profile is $\boldsymbol{B} = (B_1, B_2, B_3, B_4, B_5)$ with $B_1 = B_2 = (1,1,1)$, $B_3 = B_5 = (1,0,0)$, and $B_4 = (0,1,0)$. We have $F_{0.5}(\boldsymbol{B}) = (1,1,0)$ and $F_{0.67}(\boldsymbol{B}) = (1,0,0)$. Writing p_1 rather than *Warm*, and so forth, the first integrity constraint becomes $p_1 \wedge p_2 \rightarrow p_3$, which can equivalently be written as the 3-clause $\neg p_1 \vee \neg p_2 \vee p_3$. Of the $2^3 = 8$ possible ballots, $(1,1,0)$ is the only one not satisfying this constraint, i.e., $(1,1,0) \notin \text{Mod}(\neg p_1 \vee \neg p_2 \vee p_3)$. Thus, $F_{0.5}(\boldsymbol{B})$ is not rational, while $F_{0.67}(\boldsymbol{B})$ is. $\square$

An aggregation rule F is called *collectively rational* w.r.t. an integrity constraint Γ if and only if we get $F(\boldsymbol{B}) \in \text{Mod}(\Gamma)$ for every profile $\boldsymbol{B} = (B_1, \ldots, B_n)$ with $B_i \in \text{Mod}(\Gamma)$ for all $i \in \mathcal{N}$ (in other words, if rationality of all ballots in the profile implies rationality of the outcome). Thus, our example, for instance, proves that the rule $F_{0.5}$ is not collectively rational (while it is not conclusive on $F_{0.67}$).

2.2 Known Results

When Γ is a positive clause, to obtain a violation we must reject all of the literals in Γ in the outcome, i.e., this will only happen for relatively high quotas. Thus, low quotas are likely to preserve rationality for integrity constraints that are positive, and by a similar argument, high quotas are likely to ensure rationality for negative integrity constraints. By a known result [7, Corollary 24], we can fully characterise the class of uniform quota rules that are collectively rational for certain integrity constraints:[1]

Proposition 1. *A uniform quota rule F_q with quota $q \in [0,1]$ is collectively rational w.r.t.:*

(i) *an integrity constraint that is a k-pclause if and only if $q \cdot n \leqslant \lceil \frac{n}{k} \rceil$;*
(ii) *an integrity constraint that is a k-nclause if and only if $q \cdot n > \lfloor \frac{n \cdot (k-1)}{k} \rfloor$.*

Note that in case k divides n, these bounds simplify to $q \leqslant \frac{1}{k}$ and $q > \frac{k-1}{k}$, respectively. As it is known that the set of integrity constraints w.r.t. which a given aggregation rule is collectively rational is closed under taking conjunctions [7, Lemma 3], the above bounds also apply to conjunctions of k-pclauses and conjunctions of k-nclauses, respectively. These results mirror similar results for other frameworks of judgment aggregation [4]. For integrity constraints that mix positive and negative literals, the conditions under which collective rationality can be guaranteed currently are not as well understood.

Importantly, quota rules without guaranteed collective rationality may still return rational outcomes in a significant number of cases in practice. The purpose of our experiments will be to understand how well these rules do when the assumptions of Proposition 1 are not satisfied.

3 Generation of Data for Experiments

For our experiments, we require several aggregation problems and large numbers of profiles for these aggregation problems. In this section, we describe how we have generated this data.

3.1 Drawing an Integrity Constraint

An aggregation problem is defined by a number of issues m and an integrity constraint Γ involving (at most) the propositional variables $p_1, \ldots, p_m$. We have generated integrity constraints Γ in CNF characterised by the following three parameters (besides m):

- $\ell \in \mathbb{N}$: the number of clauses in Γ (which might be 1, if Γ is simply a clause);
- $k \in \mathbb{N}$: the number of literals in each clause;
- $k^+ \leqslant k$: the number of literals per clause that are positive.

[1] Our statement of this result differs from the original [7, Corollary 24], because we represent quotas as ratios (numbers between 0 and 1), while in the original work they are represented as absolute thresholds (numbers between 0 and n).

Thus, for $k^+ = k$ we obtain k-pclauses and for $k^+ = 0$ we obtain k-nclauses. W.l.o.g., in every clause we let *the first* k^+ *literals* be the positive ones. As an example, the integrity constraint $(p_1 \vee p_5 \vee \neg p_2) \wedge (p_5 \vee p_3 \vee \neg p_1)$ could have been generated by the parameters $m = 5$, $\ell = 2$, $k = 3$, and $k^+ = 2$.

The parameters fully determine an integrity constraint, except for the identity of the propositional variable to occur in each of the k positions in a clause. We generate constraints by drawing, for each clause, a permutation of $(p_1, \ldots, p_m)$ from the uniform probability distribution over all such permutations, to then instantiate the clause with the first k elements of that permutation. This ensures that all k-clauses are well-formed in the sense of not containing the same variable more than once.

Of course, for some settings of parameters, there only exists a single integrity constraint. For example, modulo reordering of disjuncts (i.e., modulo logical equivalence), there is only one 4-pclause for 4 issues, namely $p_1 \vee p_2 \vee p_3 \vee p_4$ (here the parameters are $\ell = 1$, $k = k^+ = 4$).

3.2 Drawing a Rational Profile

Suppose we have fixed an integrity constraint Γ, as well as n (the number of agents) and m (the number of issues). We assume that there exists an objectively "correct" ballot $B^\star = (b_1^\star, \ldots, b_m^\star)$, which we draw from the uniform probability distribution over $\mathrm{Mod}(\Gamma)$, the set of all rational ballots. We further assume that each agent wants to report the "correct" ballot, but that she might make mistakes. Specifically, we assume she proceeds as follows. Fix some probability $p \geqslant 0.5$, representing the *observability* of issues for our agent. For every issue j, she correctly reproduces $b_j^\star$ with probability p. Of course, that way she might end up with a ballot that is not rational. If that happens, she throws away her ballot and tries again from scratch. Thus, to obtain her own ballot $B = (b_1, \ldots, b_m)$, she executes this algorithm:

repeat
  for $j = 1, \ldots, m$ **do**
    $b_j := \begin{cases} b_j^\star & \text{with probability } p \\ 1 - b_j^\star & \text{with probability } 1 - p \end{cases}$
  end for
until $B \in \mathrm{Mod}(\Gamma)$

For $p = 0.5$, this reduces to the *uniform distribution* over $\mathrm{Mod}(\Gamma)$, corresponding to the *impartial culture assumption* in voting theory [12]. For $p = 1.0$, every agent perfectly reproduces $B^\star$.

In practice, generating profiles in this manner would be too time-consuming. Instead, we first compute the corresponding probability distribution $\Delta^p_{\Gamma,B^\star}$ over rational ballots (parametrised by Γ, $B^\star$, and p) and then draw ballots from that distribution. Let B be a rational ballot in $\mathrm{Mod}(\Gamma)$ that disagrees with $B^\star$ on k issues. That is, k is the *Hamming distance* $H(B, B^\star) := \#\{j \in \mathcal{I} \mid b_j \neq b_j^\star\}$ between B and $B^\star$. What should be the probability $P(B)$ of drawing B under $\Delta^p_{\Gamma,B^\star}$? If we omit the rationality check at the end, then that probability is easily seen to be $P'(k) := p^{m-k} \cdot (1-p)^k$. But $P(B)$ is greater than that, as we might first draw an irrational ballot and then get a second chance to draw B. The probability of drawing one of the ballots that pass the rationality check is $P_{\text{rat}} := \sum_{B' \in \mathrm{Mod}(\Gamma)} P'(H(B', B^\star))$. That is, to compute P_{rat} we go through all rational ballots B', compute for each of them their Hamming distance to $B^\star$, and then add the probability of drawing some ballot of that distance to $B^\star$ using our first formula. The probability of drawing B then is $P(B) = P'(k)/P_{\text{rat}}$, which defines $\Delta^p_{\Gamma,B^\star}$.

To summarise, given an integrity constraint Γ and an observability p, we generate a rational profile as follows. We first draw $B^\star$ from the uniform probability distribution over $\mathrm{Mod}(\Gamma)$. We then draw n individual ballots from the probability distribution $\Delta^p_{\Gamma,B^\star}$ defined above.

4 Empirical Results

In this section, we report on a series of experiments, where we have tested how frequently quota rules with different quotas produce rational outcomes on profiles drawn from the kinds of distributions described in Section 3. A single run in an experiment consists of the application of an aggregation rule F to a rational profile for some integrity constraint. For a given experiment, consisting of a set of such runs generated using a given set of parameters, let us call the *rationality ratio* (RR) of aggregation rule F the number of runs in the set where F returns a rational outcome, divided by the total number of runs.

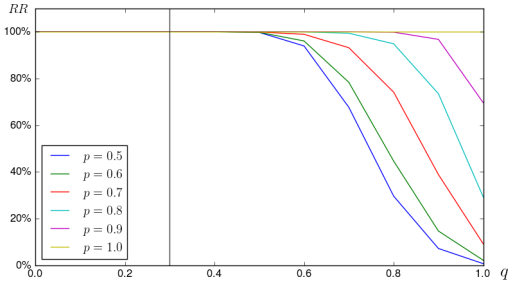
We have run each of our experiments for every quota $q \in \{0, 0.1, 0.2, \ldots, 1\}$. In the graphs, the quota q is shown on the x-axis and the average RR (in percent) is shown on the y-axis.

4.1 Effect of Observability

In our first experiment we investigate the effect of the probability p (the observability of issues for the agents) on the RR in our generative model. We do so for integrity constraints consisting either of a single positive clause or a single negative clause.

Experiments were performed for $n = 10$ and $m = 4$ on 4-pclauses and 4-nclauses (i.e., $\ell = 1$ and $k = 4$), with p ranging from 0.5 to 1.0, in steps of 0.1. Recall that $p = 0.5$ corresponds to a uniform distribution over rational ballots and $p = 1.0$ corresponds to perfect agreement amongst the agents. Note that there exist only a single 4-pclause and a single 4-nclause for this setting (modulo logical equivalence). We have run each experiment 50,000 times by generating 50,000 profiles (using the method of Section 3.2). The results are shown in Figure 1. We have repeated the same experiment for $m = k = 5$ (not shown here) to verify that small changes in the experimental setup do not have a significant impact on the results.

(a) one 4-pclause for $n = 10$ and $m = 4$



(b) one 4-nclause for $n = 10$ and $m = 4$

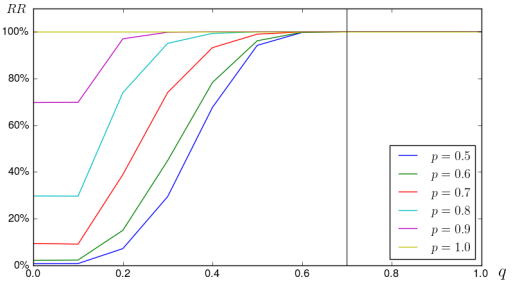


Figure 1: Effect of observability (p) of issues for 4-pclauses and 4-nclauses

First, we find that the RR increases with the observability p (i.e., as noise in the profile decreases). This is not surprising. For example, for $p = 1.0$ every agent reports the same ballot, so any quota rule will copy that ballot and be rational.

Second, we find that for positive clauses, the RR decreases as the quota increases (and *vice versa* for negative clauses). This also is not surprising (as discussed at the start of Section 2.2). But, third, we find that the rules perform much better than what could have been expected given Proposition 1 alone. The vertical lines in Figure 1 indicate the bounds for guaranteed collective rationality implied by Proposition 1. For 4-pclauses, for instance, we know that any quota equal to at most $\lceil \frac{10}{4} \rceil / 10 = 0.3$ guarantees rationality. But in fact, we also get a RR of (almost) 100% for quotas up to 0.5, even for maximal noise in the input (i.e., for observability $p = 0.5$). This can be explained by our generative model, which entails that the probability of a given issue having value 1 is just over 50% (see Section 4.2 for a precise statement of this point). Thus, the probability that, for each of the 4 issues, more than half of the agents pick a 0 becomes vanishingly small (albeit not 0).

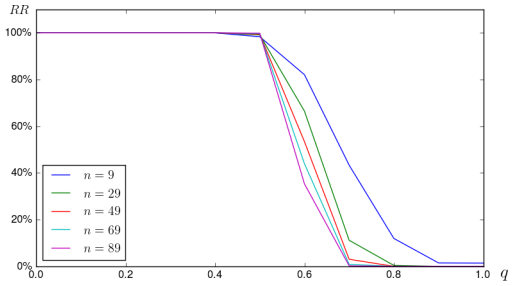
In summary, particularly but not only for groups of agents that are reasonably well aligned in their judgments, we can expect rational outcomes for a much wider range of quotas than those that guarantee collective rationality by Proposition 1.

4.2 Effect of the Number of Agents

In our second experiment, we investigate how the number of agents influences the RR. We focus on positive clauses and two specific values for p (observability), one corresponding to completely random input ($p = 0.5$) and one modelling moderately high agreement amongst agents ($p = 0.8$).

We again set $m = k = 4$ and $\ell = 1$ (i.e., we work with a single 4-pclause for 4 issues) and we let n vary from 9 to 89, in steps of 20. Note that Proposition 1 guarantees collective rationality only for quotas below roughly $\frac{1}{k} = 0.25$ for these parameters (the exact bound depends on n, and more specifically on how far off an integer $\frac{n}{k}$ is). We again have run each individual experiment 50,000 times. The results are shown in Figure 2. The effects observed are similar for even numbers of agents (not shown), but somewhat more clearly pronounced for odd n. Results for negative clauses (not shown) are analogous (in the sense in which they were for our first experiment).

(a) one 4-pclause for $m = 4$ and $p = 0.5$



(b) one 4-pclause for $m = 4$ and $p = 0.8$

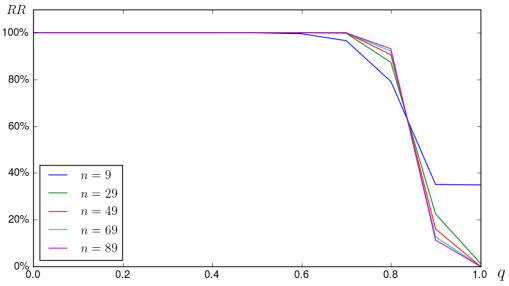


Figure 2: Effect of the number of agents (n) for 4-pclauses with $p = 0.5$ and $p = 0.8$

First, note that, as expected, the two graphs for $n = 9$ are almost identical to the corresponding graphs for $n = 10$ in the lefthand part of Figure 1. The small anomaly for $p = 0.8$ is due to the fact that for 9 agents the quota rules with $q = 1$ and $q = 0.9$ are in fact identical.

Second, we observe that the abruptness with which the RR decreases for increasing quota becomes more pronounced as we increase the number of agents n. This may be best understood by considering what happens for $n \to \infty$, given our generative model. For $\Gamma = (p_1 \vee p_2 \vee p_3 \vee p_4)$, the only irrational ballot is $(0, 0, 0, 0)$, leaving $2^4 - 1 = 15$ rational ballots. Thus, if we draw ballots from the uniform probability distribution over Mod(Γ), as we do for $p = 0.5$, the probability of a given issue getting value 1 is $\frac{8}{15} \approx 0.53$. Thus, by the Law of Large Numbers, for $n \to \infty$, we can expect for every issue j almost exactly 53% of the agents picking a 1. Thus, for any quota below 53% the collective choice will be 1, and for any quota above 53% it will be 0. As this is so for all issues, we get the rational outcome $(1, 1, 1, 1)$ for small quotas and the irrational outcomes $(0, 0, 0, 0)$ for high quotas. This situation is similar for other values of p. For small values of n, we still get this general effect as far as very low and very high quotas are concerned, but things change more smoothly in between. This also explains why the lines for different n cross at some point. This effect is more pronounced for $p = 0.8$, but it is also present for $p = 0.5$.

Thus, as before, we see that in practice we can work safely with significantly higher quotas than the known theoretical results might suggest. Furthermore, for moderate numbers of agents performance decreases only relatively slowly as we increase the quota, while for large numbers of agents there is a fairly abrupt change, so choosing a suitable quota becomes more critical.

4.3 Effect of Using Mixed Clauses

To be able to compare our empirical findings with the theoretical bounds offered by Proposition 1, in our first two experiments we have restricted attention to very simple integrity constraints, either a single positive clause or a single negative clause. However, in practice integrity constraints can be expected to be much more complex. To address this, in our third experiment we work with constraints consisting of more than one clause, each of which mixes positive and negative literals.

Experiments were performed for $n = 10$ and $m = 4$, for both $p = 0.5$ and $p = 0.8$. In each experiment, the integrity constraint consists of two clauses ($\ell = 2$). The variables varied are k (the length of each clause) and k^+ (the number of positive literals per clause). For each individual experiment, we have generated 5 different integrity constraints (using the method of Section 3.1), to prevent results from being heavily dependent on a specific choice of constraint. Then, for each of these integrity constraints (and for each quota $q \in \{0, 0.1, 0.2, \ldots, 1\}$), we have generated 10,000 rational profiles (thus, we again have a total of 50,000 runs per quota). The results are shown in Figure 3.

First, note that in case either all literals are positive or all literals are negative, we obtain graphs very similar to those in Figure 1, i.e., the effect of now having two rather than just one clause as well as the effect of having shorter clauses is, in itself, not highly significant. We also again see the effect discussed before, of increased observability ($p = 0.8$ rather than $p = 0.5$) improving results.

Second, in case clauses are truly mixed, i.e., neither purely positive nor purely negative, we see a very marked increase of the RR. We can explain this effect as follows. If q is relatively low, then issues tend to get accepted by F_q, so positive literals in the integrity constraint are likely to be satisfied (recall

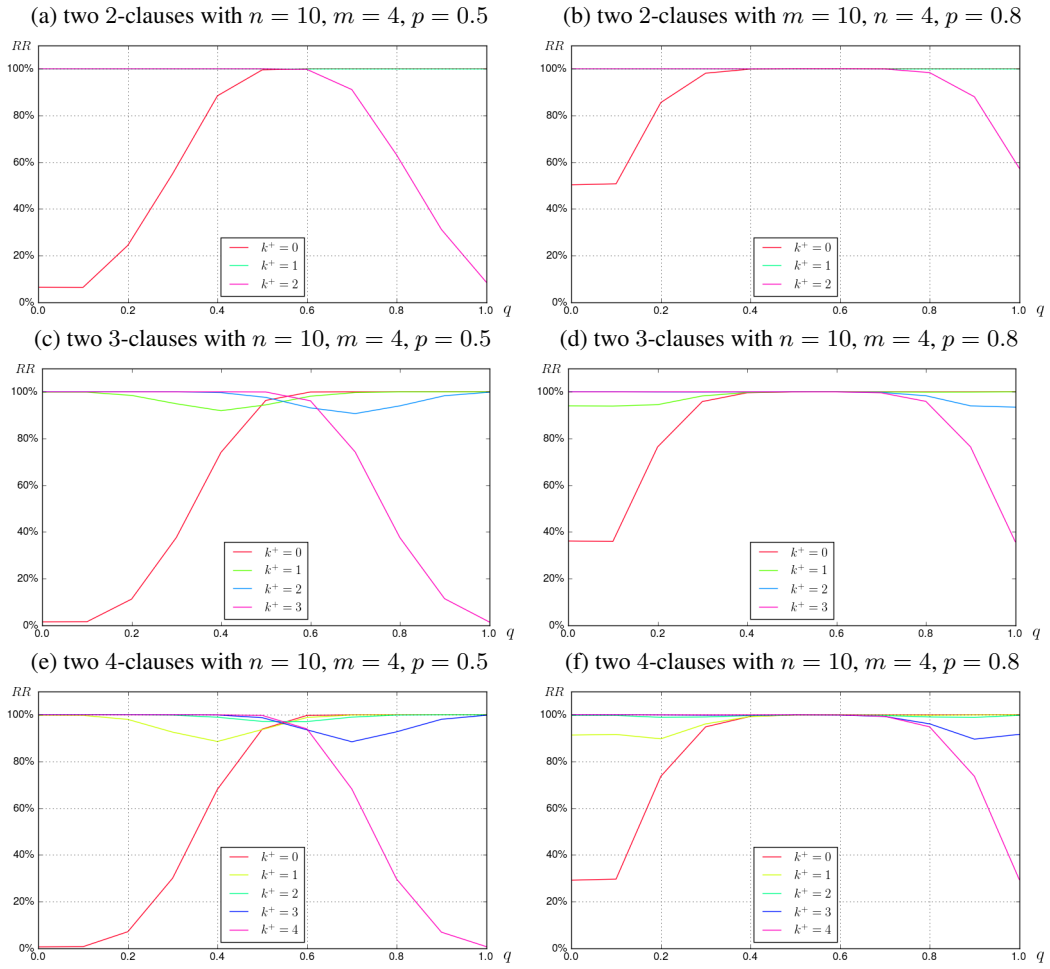


Figure 3: Effect of length of clauses ($k = 2, 3, 4$) and number of positive literals ($k^+ = 0, \ldots, k$)

that satisfying just one literal in a clause is enough to satisfy the clause). If q is relatively high, by the same kind of argument, the negative literals are likely to be satisfied. Thus, if a clause has both positive and negative literals, then chances are good that at least one of these two ways of satisfying the clause is triggered, whatever the quota. This analysis also explains why for the 4-clauses the case of $k^+ = 2$ (i.e., the case that balances positive and negative literals the most) works best.

Third, we can see some evidence for performance being better with shorter clauses, although it is difficult to isolate these effects from the effect caused by varying the proportion of positive literals. We can offer some speculative explanation here. If clauses are shorter, the set of rational ballots is smaller. Thus, shorter clauses imply increased cohesion in the profile, which is likely to improve the RR. At the same time, having fewer ballots that are rational also means that it will be harder to satisfy rationality in the outcome, i.e., here we have an effect pulling in the opposite direction. Our experimental findings thus suggest that, in practice, the first effect is somewhat stronger than the second.

Overall, these are very positive results. For the most realistic scenario, namely integrity constraints being conjunctions of several mixed clauses, albeit being harder to analyse than the more clearly structured earlier cases, we in fact obtain the best performance in terms of RR.

5 Conclusion

We have evaluated the rationality ratio achieved by quota rules for synthetically generated profiles of judgments for aggregation problems characterised by different integrity constraints and we have compared the results obtained to known theoretical bounds for collective rationality guarantees. This com-

parison indicates that quotas can be set more freely in practice than those theoretical results would suggest, especially in cases where the agents tend to agree with each other, i.e., in cases where the Hamming distances between their judgments are small. This happens, for instance, when each agent reports a noisy version of particular judgment, as might be the case when several experts are asked to offer judgment, or when the agents are voters with similar political convictions. As we have seen, when the integrity constraint consists of several clauses that mix positive and negative literals, the rationality ratio tends to be higher than for the simple cases of purely positive or purely negative clauses covered by the theoretical results. This is an encouraging result, as real-life integrity constraints will more often than not exhibit this mixed structure.

Our approach to generating data for these experiments, while capturing some intuitions about typical features of realistic judgment profiles, is still relatively simplistic. In future work, this should be complemented with experiments using real-world data. This is challenging, as real-world data of this kind is not yet readily available, at least not in domains where collective rationality matters.[2] An alternative direction therefore may be to come up with richer generative models, e.g., by defining a Bayesian Network [2] over issues. This can make the generation process more natural, since voters typically hold beliefs that influence their judgments on multiple issues. For example, an agent who believes that emissions from cars increase global warming will be more likely to believe that the number of cars should be reduced. We expect that the Bayesian Network approach will lead to more positive results than predicted by known theoretical results as well, because the Hamming distance between agents that accept a specific issue will be small, leading to relative uniformity in profiles.

References

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

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[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] 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.

[6] U. Grandi and U. Endriss. Binary aggregation with integrity constraints. In *Proc. 22nd International Joint Conference on Artificial Intelligence (IJCAI-2011)*, 2011.

[7] U. Grandi and U. Endriss. Lifting integrity constraints in binary aggregation. *Artificial Intelligence*, 199–200:45–66, 2013.

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

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

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

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

[12] M. Regenwetter, B. Grofman, A. A. J. Marley, and I. Tsetlin. *Behavioral Social Choice: Probabilistic Models, Statistical Inference, and Applications*. Cambridge University Press, 2006.

[13] Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2009.

[14] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*. ACL, 2008.

[15] M. Wooldridge. *An Introduction to Multiagent Systems*. John Wiley and Sons, 2nd edition, 2009.

[2] There is empirical data for judgment aggregation, particularly data collected in crowdsourcing exercises, but only for problems without integrity constraints. The data on annotations of linguistic corpora collected by Snow et al. [14] is an example.