Social Choice and Welfare (2019) 53:415–442
https://doi.org/10.1007/s00355-019-01191-9

ORIGINAL PAPER

Strategyproof judgment aggregation under partial information

Zoi Terzopoulou[1] · Ulle Endriss[1]

Received: 9 November 2017 / Accepted: 17 April 2019 / Published online: 20 April 2019

Abstract
We introduce a model of judgment aggregation in which individuals do not necessarily have full information regarding the judgments held by their peers. This intuitively limits an individual's ability to strategically manipulate the aggregation process. Our results confirm this basic intuition. Specifically, we show that known impossibility results concerning the existence of reasonable strategyproof judgment aggregation rules break down once we abandon the classical assumption of full information. For instance, the simple plurality rule is strategyproof in case individuals do not have any information about their peers, while the well-known premise-based rule can be rendered strategyproof by withholding only a negligible amount of information.

1 Introduction

The framework of logic-based judgment aggregation introduced by List and Pettit (2002) provides a rich environment in which to study collective decision making. It has been found useful by researchers in Legal Theory, Philosophy, Logic, Mathematics, Economics, Political Science, Computer Science, and Artificial Intelligence (alike see, e.g. List and Puppe 2009; List 2012; Grossi and Pigozzi 2014; Endriss 2016). While there now is a substantial and significant body of literature on a variety of topics in judgment aggregation, the analysis of the incentives of individuals to misrepresent their own judgments has only received limited attention to date. One aspect in particular so far has been ignored entirely, namely the fact that in practice an individual considering to manipulate the outcome of a judgment aggregation rule typically will not have full information regarding the judgments to be submitted by her peers, which intuitively constrains her own ability to manipulate successfully. To address this shortcoming,

The authors thank the associate editor and the two reviewers for their valuable suggestions.

✉ Zoi Terzopoulou
z.terzopoulou@uva.nl

[1] Institute for Logic, Language and Computation (ILLC), University of Amsterdam, Postbus 94242, 1090 GE Amsterdam, The Netherlands

in this paper, we propose a model of strategic manipulation in judgment aggregation under partial information.

Example. To illustrate the basic idea, we begin with an example, inspired by Bovens and Rabinowicz (2006). Suppose a committee of five professors—Alice, Bob, Carol, Deniz, and Enrique—has to decide whether one of their junior colleagues should receive tenure. Regulations stipulate that the candidate needs to be found to perform at an excellent level with regards to research, teaching, and service to the profession. They also stipulate that on each of these issues the committee should decide by majority, and tenure should be granted if and only if there is a majority for each of the three issues. The private views of the committee members are shown in the following table:

Committee	Research?	Teaching?	Service?	Tenure?
Alice	Yes	No	Yes	No
Bob	Yes	Yes	Yes	Yes
Carol	Yes	Yes	Yes	Yes
Deniz	No	No	No	No
Enrique	No	Yes	Yes	No

Let us refer to the issues of excellence in research, teaching, and service as the *premises* and to the issue of whether tenure will be granted as the *conclusion*. Thus, the committee is using the well-known *premise-based rule* (Pettit 2001; Dietrich and Mongin 2010). If all committee members provide truthful judgments, then tenure will be granted, as there will be a majority on each of the three premises.

Now consider Alice, who would not be entirely happy with this outcome. If she—somehow—knows the others will vote as shown in the table, then she can lie and claim that she believes the candidate's performance on research to be insufficient. In that case, the majority decision on research would come out negative and tenure would not be granted. Thus, given full information and assuming Alice only cares about the conclusion, she has an incentive to manipulate. In fact, this does not change if she has less information. Even if she has no information at all about the expected judgments of her colleagues, she could simply vote "No" on all three premises, which would be a safe strategy that would avoid the candidate getting tenure whenever Alice's judgment is pivotal on any of the premises she truthfully accepts, and it would simply not change the outcome on the conclusion in all other cases.

But for the actual scenario above, a lie by Alice on the candidate's performance regarding service would be an ineffective lie, as the majority in favour is too strong to begin with. Now suppose Alice cares not only about the conclusion: the outcome on the conclusion is what is most important to her but, all else being equal, she would rather have the collective judgment on the premises recorded in the official minutes of the meeting to be as close as possible to her own truthful judgments. Thus, if it does not affect the conclusion, she would not want to lie about a given premise. Hence, if Alice does not have full information and is unsure whether she is pivotal on the issues of research and service, then she cannot safely manipulate.

Related work The idea of modelling strategic individuals who may manipulate by misrepresenting their truthful judgments was introduced into the judgment aggregation literature by Dietrich and List (2007c).[1] Adopting an axiomatic perspective, Dietrich and List characterise the family of all strategyproof judgment aggregation rules, i.e., all rules that are immune to strategic manipulation, but also show that all rules that have certain desirable properties must be manipulable. These—largely negative—results presuppose full information on the part of any potential manipulator.

In search for more positive news, one route to take is to impose *domain restrictions*, i.e., to limit the range of judgments that a group can submit. For example, if the group's judgments are always *unidimensionally aligned*,[2] then the *majority rule*—which is known to be strategyproof (Dietrich and List 2007c)—does not suffer from the familiar problem of sometimes returning inconsistent outcomes (List 2003). A second approach is to look for aggregation rules for which strategic manipulation may be mathematically (hence theoretically) possible, but computationally (hence practically) intractable. Such *computational barriers* against strategic manipulation in judgment aggregation were first considered by Endriss et al. (2012). For example, the problem of deciding whether the premise-based rule can be manipulated successfully in a given profile is NP-complete. Further results of this kind have been obtained by Baumeister et al. (2015) and de Haan (2017).

But arguably the most natural approach towards containing strategic manipulation, namely *informational barriers* against manipulation, so far has been neglected in research on judgment aggregation. The situation is somewhat different in other areas of social choice theory. While voting and preference aggregation suffer from similarly negative results in the general case—starting with the seminal Gibbard-Sattherwhate Theorem (Gibbard 1973; Satterthwaite 1975)—and while domain restrictions (Gaertner 2001) and computational barriers to manipulation (Conitzer and Walsh 2016) have been investigated extensively in this domain as well, in the context of voting there also have been several attempts at capturing the notion the partiality of the information available to a manipulator in an election.

For instance, Osborne and Rubinstein (2003) propose a model under which every voter knows the preferences of a small *sample* of the electorate and believes that this sample is representative. Chopra et al. (2004) work with a directed graph, called the *knowledge graph*, where voter i is taken to know the preferences of voter j if there is an edge from node i to node j in the graph. Conitzer et al. (2011) investigate a more general setting, where the *set of possible preference profiles* $\mathcal{W}_i$ that a voter i deems possible is given explicitly. This model is developed further by Reijngoud and

[1] Somewhat further removed from the concerns of this paper, other authors have investigated other forms of strategic behaviour in judgment aggregation, notably *group manipulation* (Botan et al. 2016), *bribery* (Baumeister et al. 2015), and *agenda setting* (Dietrich 2016). Moreover, in the context of *epistemic* judgment aggregation, i.e., when a *ground truth* can be assumed to exist about the issues under consideration, the strategic behaviour of partially informed individuals has been investigated by Bozbay et al. (2014). For a broader perspective, we refer to the recent survey by Baumeister et al. (2017), which specifically emphasises algorithmic considerations.

[2] The judgments submitted by a group are said to be unidimensionally aligned, if the members of that group can be lined up from left to right, such that, for every issue upon which judgments are expressed, all the individuals that have the same opinion are either all on the left or all on the right side of those that disagree with them (List 2003).

Endriss (2012), who assume that the information available to each voter is induced by some *opinion poll*. This model, also employed by Endriss et al. (2016), is the closest to the one we are going to develop in this paper. Finally, in a related model explored by Meir et al. (2014), each voter is taken to consider possible the set of preference profiles in some *neighbourhood* of the true profile. Most of these authors, like us in this paper, assume that individuals are extremely risk-averse, in the sense that they will only consider manipulating if, given the information available to them, they consider it *certain* that the outcome will not be worse than if they vote truthfully and they consider it *possible* that the outcome will be strictly better.

Our contribution In this paper, we introduce a model of strategic manipulation in judgment aggregation under partial information and show that some—but not all—known negative results concerning the existence of reasonable strategyproof judgment aggregation rules break down once we abandon the classical assumption of full information. Concretely, the simple *plurality rule*, which selects the overall judgment chosen by the largest number of individuals and which fails to be strategyproof under full information (Dietrich and List 2007c)—is immune to strategic manipulation if individuals do not have any information about the judgments of others at all. While assuming *zero information* is a strong assumption, another result shows that we can always find an aggregation rule that fails to be strategyproof under full information but is strategyproof under partial (but nonzero) information.

Our results regarding one of the most important judgment aggregation rules used in practice, the premise-based rule, heavily depend on the assumptions we wish to make regarding the preferences of the individuals, which—together with the information they have access to and the aggregation rule in use—determine their incentives to manipulate. Specifically, we distinguish whether individuals only care about certain issues or whether they care about different issues to different degrees. Our results show that (i) when individuals care much more about conclusions than premises, then the premise-based rule can be rendered strategyproof by withholding only a negligible amount of information and that (ii) when individuals care equally about all issues, then the premise-based rule is strategyproof under both full and partial information. These results provide an interesting contrast with a result due to Dietrich and List (2007c), which shows that when individuals care only about the conclusion, then the premise-based rule fails to be strategyproof (this also is true under both full and partial information).

Paper overview The remainder of this paper is organised as follows. In Sect. 2, we summarise some of the fundamentals of judgment aggregation and fix our notation. In Sect. 3, we first recall relevant definitions and results due to Dietrich and List (2007c) and then introduce our model of strategic manipulation under partial information. Several results highlighting the differences between full and partial information are given in Sect. 4, while Sect. 5 focuses on results regarding the premise-based rule.[3] Section 6 concludes. Proofs of all technical results have been relegated to the "Appendix".

[3] Further results pertaining to our model may be found in the Master's thesis of the first author (Terzopoulou 2017).

2 The framework of judgment aggregation

In this section, we recall the standard model of judgment aggregation (List and Puppe 2009; List 2012; Grossi and Pigozzi 2014; Endriss 2016), originally introduced by List and Pettit (2002).

Consider a finite set of *individuals* $N = \{1, 2, \ldots, n\}$, with $n \geq 2$, that constitute a group whose judgments are to be aggregated into one collective decision. The judgments of the individuals are represented as formulas in classical propositional logic.[4]

2.1 Agendas

The domain of decision making is an *agenda*, a nonempty set of formulas of the form $\Phi = \Phi^+ \cup \{\neg\varphi : \varphi \in \Phi^+\}$, where the *pre-agenda* Φ^+ consists of non-negated formulas only.

Several restrictions can be imposed on the structure of an agenda, in order to better capture the essence of specific aggregation situations.[5] For instance, a *conjunctive agenda* Φ consists of a set of *premises* Φ^p and a single *conclusion* (together with its negation). The latter is understood to be satisfied if and only if all premises are (Dietrich and List 2007c). The example given in the introduction uses an instance of a conjunctive agenda, as do many examples discussed in the literature (see, e.g., List and Pettit 2002; Hartmann and Sprenger 2012). Formally, the pre-agenda of a conjunctive agenda is of the form $\Phi^+ = \{p_1, \ldots, p_k, c\}$, with the p_j being propositional variables and $c = (p_1 \wedge \cdots \wedge p_k)$. Analogously, in a *disjunctive agenda* the conclusion is equivalent to the disjunction of all the (non-negated) premises. Conjunctive and disjunctive agendas appear in situations in which a final decision has to be made on a conclusion, but the reasons that lead to that choice, described by the premises, are also important.

A more general class of agendas, which includes conjunctive and disjunctive ones, is that of the *path-connected* agendas (Dietrich and List 2007a), related to the concept of *total-blockedness* (Nehring and Puppe 2007). An agenda is path-connected if any two of its formulas are logically connected with each other, either directly or indirectly, via a sequence of conditional logical entailments. Formula φ *conditionally entails* formula ψ if $\{\varphi, \neg\psi\} \cup \Psi$ is logically inconsistent for some $\Psi \subseteq \Phi$ logically consistent with φ and with $\neg\psi$. The agenda Φ is path-connected if for all propositions $\varphi, \psi \in \Phi$ that are neither tautologies nor contradictions, there is a sequence of propositions $\varphi_1, \varphi_2, \ldots, \varphi_k \in \Phi$ with $\varphi = \varphi_1$ and $\psi = \varphi_k$ such that φ_{i-1} conditionally entails φ_i, for every $i \in \{2, \ldots k\}$. Many standard and interesting agendas are path-connected, but note that conjunctive agendas and disjunctive agendas are not path-connected: in a conjunctive agenda there are no conditional entailments from non-negated towards negated formulas.

[4] An extended model of judgment aggregation which captures propositions expressed in richer logical languages, such as predicate logic, modal logic, and multivalued or fuzzy logic has been developed by Dietrich (2007).

[5] Another reason to focus on agendas with certain properties is to ensure good behaviour of aggregation rules, particularly the logical consistency of the collective outcome (see, e.g., Nehring and Puppe 2007; Dietrich and List 2007a; Endriss et al. 2012).

2.2 Aggregating individual judgments

Each individual i has a *judgment set* $J_i \subseteq \Phi$, the set of formulas she accepts. We assume that all individual judgment sets are *consistent*, i.e., logically consistent sets of formulas, and *complete*, i.e., $\varphi \in J_i$ or $\neg\varphi \in J_i$ for every $\varphi \in \Phi^+$.[6] The set of all consistent and complete subsets of the agenda is denoted as $\mathcal{J}(\Phi)$. A *profile* $\boldsymbol{J} = (J_1, \ldots, J_n) \in \mathcal{J}(\Phi)^n$ is a vector of all the individual judgment sets, and $\boldsymbol{J}_{-i}$ stands for the partial profile of judgments of the whole group besides individual i. We denote with $N_\varphi^{\boldsymbol{J}}$ the set $\{i : \varphi \in J_i\}$ of individuals who accept formula φ in profile $\boldsymbol{J}$. We write $\overline{J}$ for the complement $\Phi \setminus J$ of any given judgment set $J \subseteq \Phi$. Furthermore, we say that the judgment set J *agrees with* the judgment set J' on formula φ whenever $\varphi \in J \cap J'$ or $\varphi \in \overline{J} \cap \overline{J'}$, and that J *disagrees with* J' otherwise.

There are various methods to aggregate the judgments of a group, which lead to different collective outcomes. An *aggregation rule* F is a function that maps every profile of judgments $\boldsymbol{J} \in \mathcal{J}(\Phi)^n$ to a nonempty set of nonempty collective judgments, i.e., to a nonempty subset of $2^\Phi \setminus \emptyset$, where 2^Φ is the powerset of Φ. Thus, there may be a tie between several "best" judgment sets and these judgment sets need not be consistent or complete. When $F(\boldsymbol{J})$ is always a singleton, that is, when $F : \mathcal{J}(\Phi)^n \to 2^\Phi \setminus \emptyset$, the rule F is called *resolute*. In practice, the aim of an aggregation rule is to provide us with an answer about what the collective decision of the individuals is, or should be. Hence, resoluteness is essential, and we can guarantee it by considering a *lexicographic tie-breaking rule* to resolve the ties between the suggested collective opinions.[7]

2.3 Specific aggregation rules

A straightforward judgment aggregation rule is the *majority rule*, which accepts a formula in the agenda if and only if at least half of the individuals accept it. The *quota rules* generalise this idea. According to them, a formula φ is part of the collective decision if and only if at least a certain proportion of the individuals (meeting the relevant *quota* $q_\varphi \in [0, n+1]$) agrees with accepting φ. Formally, the quota rule F^q is such that, for any profile $\boldsymbol{J}$:

$$\varphi \in F^q(\boldsymbol{J}) \quad \text{if and only if} \quad \left|N_\varphi^{\boldsymbol{J}}\right| \geq q_\varphi.$$

Unfortunately, when a quota rule is used, the collective outcome may end up being logically inconsistent. This is the case, for instance, for the example given in the intro-

[6] For a discussion of the relaxation of the completeness assumption, we refer to the work of Gärdenfors (2006), Dietrich and List (2008), and Terzopoulou et al. (2018).

[7] An alternative technique of breaking ties could exploit *random tie-breaking*. However, we restrict attention to lexicographic tie-breaking orders. One reason is that breaking ties with the help of a fixed linear order satisfies the *independence of irrelevant alternatives principle* (Ray 1973). The independence of irrelevant alternatives principle, also known as Sen's property α (Sen 1969, 1970), states that if an alternative J is chosen from a set S, and J is also an element of a subset S' of S, then J must be chosen from S'. That is, eliminating some of the unchosen alternatives should not affect the selection of J. We find this condition normatively desirable as far as the tie-breaking rule is concerned.

duction (when the individuals are asked to provide judgments on research, teaching, service, and tenure), if the quota rule F_q with, say, $q_\varphi = 3$ for all φ is applied.

The most popular way to resolve this problem is to use the *premise-based rule* F^{pr} (Pettit 2001; Chapman 2002; Dietrich and List 2007b; Dietrich and Mongin 2010; Hartmann and Sprenger 2012), which we define here with regard to conjunctive agendas only. First, a collective decision is made on the premises with respect to the (strict) majority rule. Concretely, for all $p_i \in \Phi^p \cap \Phi^+$, $p_i \in F^{pr}(\boldsymbol{J})$ if $|N_p^{\boldsymbol{J}}| > \frac{n}{2}$, and $\neg p \in F^{pr}(\boldsymbol{J})$ otherwise. Then, the conclusion is accepted by the group if and only if all the premises are: $c \in F^{pr}(\boldsymbol{J})$ if $p \in F^{pr}(\boldsymbol{J})$ for all $p \in \Phi^p \cap \Phi^+$, and $\neg c \in F^{pr}(\boldsymbol{J})$ otherwise. Since $c = (p_1 \wedge \cdots \wedge p_k)$, a consistent outcome is then guaranteed. The definition for disjunctive agendas is analogous.

The premise-based rule applied to conjunctive agendas has received noticeable attention by economists and philosophers, especially because of its significance in the domains of politics and law (Pettit 2001; Chapman 2002). A famous argument in favour of the premise-based way of aggregating individual judgments relates to deliberative democracy (Elster 1998), supporting the view that collective decisions on conclusions should be determined by the group's opinions on the premises.

Next, we turn to a family of aggregation rules that most will find objectionable, the *dictatorships*. Living up to its name, a dictatorship is connected to a single individual, the dictator, whose judgment is taken to be the collective judgment independently of the input profile. So, F is a dictatorship if and only if there exists an individual $i \in N$ such that $F(\boldsymbol{J}) = \{J_i\}$, for every profile $\boldsymbol{J} = (J_1, \ldots, J_n)$.

Observe that the quota rules, the premise-based rule, and the dictatorships are all resolute by construction.

A dictatorship vacuously guarantees that the collective judgment will satisfy all the nice properties of individual judgments, like completeness and consistency. Fortunately, there are several other aggregation rules that also exhibit this advantage. Such an aggregation rule, directly inspired by voting theory (see, e.g., Zwicker 2016), is the *plurality rule*. The plurality rule $F^{p\ell}$ considers the aggregated outcome to be the judgment set(s) submitted by the largest number of individuals, i.e., for $\boldsymbol{J} = (J_1, \ldots, J_n)$ we get:

$$F^{p\ell}(\boldsymbol{J}) = \operatorname*{argmax}_{J \subseteq \Phi} |\{i \in N : J = J_i\}|.$$

The plurality rule presents certain theoretical limitations. For instance, it does not capture the internal logical structure of the judgment sets. Moreover, in settings with few individuals but many alternative judgments, it is very probable that several judgment sets receive the same amount of support, only by one individual, and hence the tie-breaking rule plays an overly important role in deciding the final outcome. Nonetheless, the plurality rule is of course widely used in political elections. Furthermore, in a different context motivated by applications to crowdsourcing, Caragiannis et al. (2014) show that an aggregation rule that they call *modal ranking*, and that is equivalent to the plurality rule (for judgment sets corresponding to rankings), is the unique one satisfying certain desirable truth-tracking properties. Later on we will see that the plurality rule also plays an important role in strategyproof judgment aggrega-

tion, since it turns out to be immune to strategic manipulation for partially informed individuals.

Finally, another rule that—like the dictatorships and the plurality rule—always selects from the judgment sets submitted by the individuals is the *average-voter rule* (Endriss and Grandi 2014). The average-voter rule F^{av} takes into account a notion of distance between judgment sets and specifies the winners to be the individual judgment sets that minimise the average distance to the elements of the profile submitted by the group. Specifically, the *Hamming-distance* of two judgment sets $J, J' \in 2^{\Phi}$ is defined as the number of formulas in Φ on which they disagree:

$$H(J, J') = |\Phi| - |J \cap J'| - |\overline{J} \cap \overline{J'}|.$$

The Hamming-distance $H(\boldsymbol{J}, J)$ of the profile $\boldsymbol{J} = (J_1, \ldots, J_n)$ and the judgment set J is the sum of the Hamming distances of all judgment sets in $\boldsymbol{J}$ and J. That is, $H(\boldsymbol{J}, J) = \sum_{i \in N} H(J_i, J)$. Then, given the profile $\boldsymbol{J}$,

$$F^{av}(\boldsymbol{J}) = \operatorname*{argmin}_{J_i \in \{J_1, \ldots, J_n\}} H(\boldsymbol{J}, J_i).$$

2.4 Properties of aggregation rules

Under a descriptive perspective, axioms provide a structured way of looking into aggregation rules, by helping us to compare them and better understand them. Under a normative perspective, axioms can guide the design of aggregation rules, as they directly reflect the properties we wish our rules to satisfy. Here, we refer to axiomatic characteristics of resolute rules only.

We are going to make use of the following axioms, all of which have been widely discussed in the literature (see, e.g., Grossi and Pigozzi 2014):

- We call an aggregation rule F *responsive* if it gives a chance to every proposition to be accepted by the group. Formally, F is responsive if, for every formula φ that is not a tautology nor a contradiction, there exist a profile $\boldsymbol{J}$ such that $\varphi \in F(\boldsymbol{J})$ and another profile $\boldsymbol{J}'$ such that $\varphi \notin F(\boldsymbol{J}')$.[8]
- *Monotonicity* prescribes that extra support for a formula $\varphi \in \Phi$ can never be damaging. Formally, F is monotonic if $\varphi \in J_i' \setminus J_i$ entails that $\varphi \in F(J_i, \boldsymbol{J}_{-i}) \Rightarrow \varphi \in F(J_i', \boldsymbol{J}_{-i})$, for all $(J_i, \boldsymbol{J}_{-i}) \in \mathcal{J}(\Phi)^n$ and $J_i' \in \mathcal{J}(\Phi)$.
- A more controversial property is *independence*, according to which each formula φ in Φ is to be treated separately by the aggregation rule F. Formally, F is independent if for all profiles $\boldsymbol{J}, \boldsymbol{J'}$, it is the case that $N_{\varphi}^{\boldsymbol{J}} = N_{\varphi}^{\boldsymbol{J'}}$ implies $\varphi \in F(\boldsymbol{J}) \Leftrightarrow \varphi \in F(\boldsymbol{J'})$. It is easy to see that, for instance, the plurality rule and the premise-based rule are not independent.

Besides enforcing axioms such as these, we can also constrain the manner in which a rule can operate by imposing conditions on the outcomes it is expected to return. In particular, an aggregation rule F is said to be *complete* (similarly *consistent*) if $F(\boldsymbol{J})$ is complete (consistent) for every $\boldsymbol{J} \in \mathcal{J}(\Phi)^n$.

[8] An alternative name of the responsiveness axiom in the literature is *nonimposition*.

3 Strategic manipulation

Let us interpret a given aggregation problem as a strategic situation, where individuals prefer certain collective decisions more than others. In this section, we introduce a model for representing such situations that emphasises the information available to each of the individuals. We start by modelling individual preferences.

3.1 Preferences

We assume that every individual i comes equipped with a preference relation $\succsim_i$, defined over all the possible collective judgment sets $J \in 2^{\Phi}$. By writing $J \succsim_i J'$, we mean that individual i wants the collective decision to be judgment J at least as much as she wants it to be judgment J'. Considering all judgment sets $J, J', J'' \in 2^{\Phi}$, we take the relation $\succsim_i$ to be *reflexive* ($J \succsim_i J$), *transitive* ($J \succsim_i J'$ and $J' \succsim_i J''$ implies $J \succsim_i J''$), and *complete* (either $J \succsim_i J'$ or $J' \succsim_i J$). Thus, we assume that individuals rank all pairs of possible outcomes relative to each other; no two outcomes are going to be incomparable.[9] Finally, we write $J \sim_i J'$ if $J \succsim_i J'$ and $J' \succsim_i J$, and we denote by $J \succ_i J'$ the strict component of $J \succsim_i J'$, i.e., the case where $J \succsim_i J'$, but not $J' \succsim_i J$.

The type of preferences that the individuals hold will play a crucial role in our analysis. So, let us reflect on some further assumptions that we can make about them. For example, in many aggregation contexts it is natural to suppose that the preferences of an individual depend on the truthful judgment set that this individual holds. Recall, for instance, the example presented in the introduction. In such a situation, it would be reasonable to assume that each individual would like the final collective decision to match her own judgment. Hence, following Dietrich and List (2007c), we restrict our study to cases where individual judgments and preferences over collective outcomes are expected to be related. A full identification of scenarios that satisfy our assumptions is an empirical problem, which certainly deserves further investigation.

A preference relation $\succsim_i$ *respects closeness* to J_i if, for any two judgment sets J and J', $J \succsim_i J'$ whenever $J \cap J_i \supseteq J' \cap J_i$. For each judgment set J_i, let $C(J_i)$ be the set of all preference relations $\succsim_i$ that respect closeness to J_i. Then, $C = \{C(J_i) : J_i \in \mathcal{J}(\Phi)\}$ is the class of *closeness-respecting preferences*. Roughly, individuals with closeness-respecting preferences rank higher the collective judgments that agree with their individual ones.

[9] The requirement of completeness of preferences has triggered lot of discussion (e.g., Jeffrey 1983), and one of the main arguments against it is directly reflected in the judgment aggregation framework. The possible collective outcomes will usually be exponentially as many as the formulas in the agenda, and the individuals have to be able to compare all of them. Nonetheless, one justification of the completeness constraint is based on our interpretation of the individuals' preferences over the collective decisions. For example, we may think of preferences expressing "conceivable" acts and not "actual" ones, in the sense that they represent the choice dispositions of the individuals (Sen 1973; Gilboa 2009). From this perspective, completeness does not imply that the individuals should be able to rank a large number of options prior to making a decision about them; instead, it may mean that they possess an intrapersonal method to rank the different judgment sets when these judgments are presented in pairs, which induces a complete ordering. An instance of a plausible such method is defined later in this paper, and is constructed via the Hamming-distance.

3.2 Full information

Along the lines of Dietrich and List (2007c), we now develop a definition of strategyproofness of an aggregation rule, relative to a given class of preferences (such as the class of all closeness-respecting preferences). However, unlike these authors, we do not distinguish between being able to affect the outcome of an aggregation rule and the incentive of doing so, but only model strategic behaviour. Like Dietrich and List, we initially assume that every individual has full information about the profile of judgments.

When does an individual have an *incentive* to submit a dishonest judgment in an aggregation problem, under the assumption that she is fully informed about the judgments of her peers? This is the question we focus on next.

Definition 1 Consider a profile of judgments $\boldsymbol{J} = (J_1, \ldots, J_n)$ and an aggregation rule F. Individual $i \in N$ with preferences $\succsim_i$ has an *incentive to manipulate* in profile $\boldsymbol{J}$ if there exists a judgment set $J_i^* \in \mathcal{J}(\Phi)$ such that $F(J_i^*, \boldsymbol{J}_{-i}) \succ_i F(J_i, \boldsymbol{J}_{-i})$.

In words, if by submitting an untruthful judgment set an individual can change the outcome into a judgment set she strictly prefers, then she has an incentive to manipulate in this manner.

Consider a function PR that assigns to each individual i and judgment set $J_i \in \mathcal{J}(\Phi)$ a non-empty set $PR(J_i)$ of reflexive, transitive and complete preference relations $\succsim_i$. Then, by a slight abuse of notation, we also denote with PR the class of preferences constructible by that function, i.e., $PR = \{PR(J_i) : J_i \in \mathcal{J}(\Phi)\}$ (an example of such a class is the class of closeness-respecting preferences).

Definition 2 The aggregation rule F is **manipulable** for the class of preferences PR if there exist a profile $\boldsymbol{J} \in \mathcal{J}(\Phi)^n$ and an individual $i \in N$ holding preferences $\succsim_i \in PR(J_i)$ such that i has an incentive to manipulate in $\boldsymbol{J}$.

Next, the notion of *strategyproofness* of an aggregation rule captures the absence of all incentives for manipulation.[10]

Definition 3 The aggregation rule F is **strategyproof** for the class of preferences PR if, for all individuals $i \in N$, all profiles $\boldsymbol{J} = (J_1, \ldots, J_n) \in \mathcal{J}(\Phi)^n$, all preference relations $\succsim_i \in PR(J_i)$, and all judgment sets $J_i^* \in \mathcal{J}(\Phi)$, $F(J_i, \boldsymbol{J}_{-i}) \succsim_i F(J_i^*, \boldsymbol{J}_{-i})$.

Thus the rule F is strategyproof for the class of preferences PR if and only if F is not manipulable for PR. This holds only because preference relations are taken to be complete. We will also refer to strategyproofness as *immunity to manipulation* and to manipulability as *susceptibility to manipulation*.

Definition 3 implies that the individuals have a noticeable *truth-bias*. That is, in case where they equally like the outcomes obtained by the truthful and possible untruthful

[10] Note that in the original terminology of Dietrich and List (2007c), *manipulability*—as opposed to *strategyproofness*—is a preference-less concept (i.e., it only depends on the individuals' truthful judgments and not on their possible preferences). In this paper though, we use this term with respect to a definition of preferences. More details on the preference-less notion of manipulability can be found in the proof of Theorem 1 in the "Appendix".

judgments, they will choose to be honest. Said differently, they will only lie if by doing so they can obtain a *strictly* better outcome. Obraztsova et al. (2013) justify this assumption by remarking that strategising can be costly for the individuals, for example in time and cognitive effort, so they would remain truthful when they cannot unilaterally affect the outcome (so when they have nothing to gain). Nevertheless, they will still try to manipulate if they can obtain a preferable result, assuming that their reward then will exceed the cost of strategising.

Dietrich and List (2007c) axiomatised the strategyproof aggregation rules, considering groups of individuals with closeness-respecting preferences that are reflexive and transitive. They proved that the strategyproof aggregation rules are exactly those that are both independent and monotonic. Next, we establish that this characterisation result remains valid for individuals whose preferences are furthermore complete.

Theorem 1 *An aggregation rule F is strategyproof for all closeness-respecting preferences that are reflexive, transitive and complete if and only if F is both independent and monotonic.*

As an immediate consequence of the result of Dietrich and List, we have that every independent and monotonic aggregation rule must be strategyproof for every subset of the class of all closeness-respecting preferences that are reflexive and transitive (and the class of all closeness-respecting preferences that are reflexive, transitive and complete obviously is such a subset). The interest in the above result thus lies in the fact that the converse is also true.

Hence, strategyproof rules exist, but they belong to a fairly narrow family of rules. The problem with rules that are independent is that they typically are not consistent and thus of limited interest in practice. For the large class of path-connected agendas, Dietrich and List (2007c) went one step further and showed that there exist *no* reasonable rules that are strategyproof:

Theorem 2 (Dietrich and List 2007c) *For a path-connected agenda Φ, an aggregation rule F is complete, consistent, responsive and strategyproof for the class of (reflexive and transitive) closeness-respecting preferences if and only if F is a dictatorship.*

3.3 Partial information

Assuming that individuals in every situation know everything about the judgments of their peers is clearly rather stringent, particularly when we consider large groups of individuals or agendas with confidential issues. In practice, the information the individuals hold in an aggregation problem may be of different types. For example, a given individual may have the information of *how many* others hold a specific judgment set J, but she may not necessarily know *which* individuals do (which is common in election polls), or she may know everyone's judgment on formula φ, but not on ψ, and so forth. Moreover, in a different setting where individuals are connected via a *social network*, an individual may know the judgments of *some* of the others (her neighbours in the network), but she may be completely uncertain about the rest.

We call $\mathcal{I}$ the set of all possible pieces of information regarding a profile of judgment sets an individual could possibly be informed about, before the final reporting

of judgments. Following Reijngoud and Endriss (2012), who introduce a similar concept in the context of voting, we define a *judgment information function* (JIF) $\pi : N \times \mathcal{J}(\Phi)^n \to \mathcal{I}$ as a function mapping individuals and profiles of judgment sets to elements of $\mathcal{I}$. Intuitively, a JIF represents the available information for every individual, given the profile of judgments of the group. To simplify notation, we write $\pi_i(\boldsymbol{J})$ for the information of individual i about profile $\boldsymbol{J}$. The following are some natural choices for $\mathcal{I}$ and the corresponding JIF π.

– *Full* The full-JIF returns precisely the input profile for every individual:

$$\pi_i(\boldsymbol{J}) = \boldsymbol{J} \quad \text{for all } i \in N \text{ and } \boldsymbol{J} \in \mathcal{J}(\Phi)^n.$$

– *All_but_**Y*** For $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, the all_but_$\boldsymbol{Y}$-JIF returns for each individual i the judgments of the rest of the group on each formula except for the formulas in $Y_i \subseteq \Phi$:

$$\pi_i(\boldsymbol{J}) = \left(N^{\boldsymbol{J}}_{\varphi}\right)_{\varphi \in \Phi \setminus Y_i} \quad \text{for all } i \in N \text{ and } \boldsymbol{J} \in \mathcal{J}(\Phi)^n.$$

– *Besides_**I*** For $\boldsymbol{I} = (I_1, \ldots, I_n)$, the besides_$\boldsymbol{I}$-JIF returns for each individual i the judgments of the other individuals besides those in $I_i \subseteq N \setminus \{i\}$:

$$\pi_i(\boldsymbol{J}) = (J_j)_{j \in N \setminus I_i} \quad \text{for all } i \in N \text{ and } \boldsymbol{J} \in \mathcal{J}(\Phi)^n.$$

– *Zero* The zero-JIF does not return any information; it just gives us a constant value:

$$\pi_i(\boldsymbol{J}) = 0 \quad \text{for all } i \in N \text{ and } \boldsymbol{J} \in \mathcal{J}(\Phi)^n.$$

Note that the full-JIF and the zero-JIF are extreme cases both of the all_but_$\boldsymbol{Y}$-JIF and of the besides_$\boldsymbol{I}$-JIF. Full information is captured when $\boldsymbol{Y} = (\emptyset, \ldots, \emptyset)$ and $\boldsymbol{I} = (\emptyset, \ldots, \emptyset)$, while we get zero information for $\boldsymbol{Y} = (\Phi, \ldots, \Phi)$ and $\boldsymbol{I} = (N \setminus \{1\}, \ldots, N \setminus \{n\})$.

Our framework also allows for the above JIFs to be combined, letting different individuals have access to different types of information. Now, having the information expressed by a JIF π and a profile of judgments $\boldsymbol{J}$, we define the set of (partial) profiles that individual i considers possible:

$$\mathcal{W}_i^{\pi, \boldsymbol{J}} = \left\{ \boldsymbol{J}'_{-i} : \pi_i(J_i, \boldsymbol{J}'_{-i}) = \pi_i(\boldsymbol{J}) \right\}.$$

That is, $\mathcal{W}_i^{\pi, \boldsymbol{J}}$ contains all the judgments of the rest of the group that are compatible with individual i's information. In the special cases where the individuals are fully informed or completely uninformed, we of course get $\mathcal{W}_i^{\text{full}, (J_i, \boldsymbol{J}_{-i})} = \{\boldsymbol{J}_{-i}\}$ and $\mathcal{W}_i^{\text{zero}, (J_i, \boldsymbol{J}_{-i})} = \mathcal{J}(\Phi)^{n-1}$ for all $i \in N$ respectively.

Coming back to the tenure-example of the introduction, suppose that Alice (individual i), before the voting process has talked with the other members of the committee

about the performance of the candidate regarding research (r) and teaching (t), so she already knows their relevant judgments on those two criteria. On the other hand, the service to the profession (s) of the candidate has only been discussed between Alice, Bob, Carol and Deniz, but not Enrique. Call π the appropriate JIF capturing the situation and suppose that the actual judgments of the five professors are as depicted in the table of the introduction. Then, the partial profiles that Alice considers possible only differ on the judgment of Enrique concerning s. Formally, it will be $\mathcal{W}_i^{\pi,\boldsymbol{J}} = \{\boldsymbol{J}_{-i}, \boldsymbol{J}'_{-i}\}$, where $\boldsymbol{J}_{-i} = (\{r,t,s\},\{r,t,s\},\{\neg r,\neg t,\neg s\},\{\neg r,t,s\})$ and $\boldsymbol{J}'_{-i} = (\{r,t,s\},\{r,t,s\},\{\neg r,\neg t,\neg s\},\{\neg r,t,\neg s\})$.

Note that we only deal with qualitative beliefs. We assume that the individuals cannot or do not want to assign any numerical value (probability) to their beliefs about the possibility of the occurrence of each scenario concerning the judgments of the group. We observe that $\mathcal{W}_i^{\pi,\boldsymbol{J}}$ satisfies the three axioms of reflexivity (REF), symmetry (SYM), and transitivity (TRA), and hence it forms an *equivalence relation*. In other words, for every individual i and judgment set J_i, the JIF π induces a partition of the set $\mathcal{J}(\Phi)^{n-1}$. Clearly, the finest partition corresponds to the full-information case and the coarsest one to zero information.[11] Formally, for all judgment sets J_i and for all (partial) profiles $\boldsymbol{J}_{-i}$, $\boldsymbol{J}^*_{-i}$, $\boldsymbol{J}^{**}_{-i}$, the following hold:

$$
\begin{aligned}
&\text{(REF)}\quad \boldsymbol{J}_{-i} \in \mathcal{W}_i^{\pi,(J_i,\boldsymbol{J}_{-i})}\\
&\text{(SYM)}\quad \boldsymbol{J}_{-i} \in \mathcal{W}_i^{\pi,(J_i,\boldsymbol{J}^*_{-i})} \text{ implies } \boldsymbol{J}^*_{-i} \in \mathcal{W}_i^{\pi,(J_i,\boldsymbol{J}_{-i})}\\
&\text{(TRA)}\quad \boldsymbol{J}_{-i} \in \mathcal{W}_i^{\pi,(J_i,\boldsymbol{J}^*_{-i})} \text{ and } \boldsymbol{J}^*_{-i} \in \mathcal{W}_i^{\pi,(J_i,\boldsymbol{J}^{**}_{-i})} \text{ imply } \boldsymbol{J}_{-i} \in \mathcal{W}_i^{\pi,(J_i,\boldsymbol{J}^{**}_{-i})},
\end{aligned}
$$

Axiom (REF) expresses that every individual always deems possible the truthful profile of judgments. Axioms (SYM) and (TRA) together state that whenever an individual considers some profile possible, then that profile would induce the same information set as her current one.

We can now refine the standard definitions of strategyproofness, accounting for individuals with incomplete information.

Definition 4 Consider an aggregation rule F, a JIF π, and a truthful profile $\boldsymbol{J} = (J_1,\ldots,J_n) \in \mathcal{J}(\Phi)^n$. Individual $i \in N$ with preferences $\succsim_i$ has an **incentive to π-manipulate** in $\boldsymbol{J}$ if there exists a judgment set $J_i^* \in \mathcal{J}(\Phi)$ such that

1. $F(J_i^*, \boldsymbol{J}'_{-i}) \succ_i F(J_i, \boldsymbol{J}'_{-i})$, for some $\boldsymbol{J}'_{-i} \in \mathcal{W}_i^{\pi,\boldsymbol{J}}$ and,
2. $F(J_i^*, \boldsymbol{J}''_{-i}) \succsim_i F(J_i, \boldsymbol{J}''_{-i})$, for all other $\boldsymbol{J}''_{-i} \in \mathcal{W}_i^{\pi,\boldsymbol{J}}$.

This means that an individual has an incentive to manipulate under the (partial) information provided by the JIF π by reporting an untruthful judgment if there is a scenario consistent with her information that will result in a more desirable collective decision for her and there is no scenario where she will be worse off than when reporting a truthful judgment. That is, we adopt a pessimistic perspective (from the individual's standpoint), according to which an individual is willing to lie only if it

[11] Chopra et al. (2004), Reijngoud and Endriss (2012), and van Ditmarsch et al. (2013) make analogous observations in the context of voting.

is totally safe to do so. Said differently, the individuals are taken to be *risk-averse*: if there is at least one possible scenario where lying induces a less desirable result, then they remain truthful.

If there is a profile $\boldsymbol{J}$ where at least one individual has an incentive to π-manipulate, then we say that the aggregation rule is π-manipulable:

Definition 5 Consider a JIF π. The aggregation rule F is **π-manipulable** for the class of preferences PR if there are a profile $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i}) \in \mathcal{J}(\Phi)^n$ and an individual $i \in N$ holding preferences $\succsim_i \in PR(J_i)$ such that i has an incentive to π-manipulate in $\boldsymbol{J}$.

The aggregation rule F is π-strategyproof for the class of preferences PR if and only if F is not π-manipulable for PR:[12]

Definition 6 Consider a JIF π. The aggregation rule F is **π-strategyproof** for the class of preferences PR if, for all individuals $i \in N$, all profiles $\boldsymbol{J} = (J_1, \ldots, J_n) \in \mathcal{J}(\Phi)^n$, all preference relations $\succsim_i \in PR(J_i)$, and all judgment sets $J_i^* \in \mathcal{J}(\Phi)$, at least one of the following conditions holds:

1. $F(J_i, \boldsymbol{J}'_{-i}) \sim_i F(J_i^*, \boldsymbol{J}'_{-i})$, for all $\boldsymbol{J}'_{-i} \in \mathcal{W}_i^{\pi,\boldsymbol{J}}$ or,
2. $F(J_i, \boldsymbol{J}''_{-i}) \succ_i F(J_i^*, \boldsymbol{J}''_{-i})$, for some $\boldsymbol{J}''_{-i} \in \mathcal{W}_i^{\pi,\boldsymbol{J}}$.

The first condition of Definition 6 brings out a further assumption concerning the *truth-bias* of the individuals. Justifying it as in the case of full information, whenever an individual cannot unilaterally change the outcome, then she chooses to be honest, so that she does not have to bear the possible cost of strategising without being able to gain any profit. The second condition of Definition 6 is related to risk-aversion. If being sincere can induce a preferable collective decision in some scenario, then an individual will not lie and risk losing that desirable outcome.[13]

Obviously, when π is the full-JIF, π-strategyproofness (π-manipulation) is equivalent to strategyproofness (manipulation) under full information. We can further understand the importance of partial information on strategyproofness as follows. Consider an individual i. If this individual possesses full information about the judgments of the rest of the group, then she will manipulate with no second thought in case she finds an untruthful judgment that makes her better off. However, finding such an insincere judgment is not sufficient to make individual i manipulate under partial information. Then, an extra condition needs to be satisfied: for all possible scenarios, the untruthful judgment should induce a result at least as good as the one induced by the individual's truthful judgment. Loosely speaking, this second condition provides an additional layer of safety against manipulation for an aggregation rule.

[12] In order to obtain Definition 6 as a contrapositive of Definition 5, we once again make use of our assumption that preference relations are complete.

[13] Note that according to Definition 3 of simple strategyproofness, the aggregation rule F is said to be strategyproof if the first condition of Definition 6 holds, taking π to be the zero-JIF. After making this observation, one could also give an alternative interpretation to the original Definition 3. That is, it may be seen as not imposing any requirements on the information that the individuals hold, but instead asking for truthfulness to be a best response to any possible judgments of the others (namely a *dominant-strategy equilibrium* rather than only a *Nash equilibrium*). For the rest of this paper though we will stick to our first interpretation of Definition 3, which is associated with full information.

4 Comparing manipulation under full and partial information

This section presents the main theoretical results of our model. We proceed with establishing an essential bridge between the framework of manipulation under full information and the richer one that incorporates settings of partial information. Subsequently, we demonstrate how partial information can escape the manipulability of aggregation rules, which haunts the standard model of full information. Throughout this section we assume that all individuals have complete and closeness-respecting preferences.

4.1 Connecting full and partial information

We start with clarifying how known results about fully informed individuals can extend to partial information cases.

Let us call a JIF π *at least as informative as* another JIF σ if for all profiles $\boldsymbol{J}$ and all individuals i, $\mathcal{W}_i^{\pi,\boldsymbol{J}} \subseteq \mathcal{W}_i^{\sigma,\boldsymbol{J}}$.[14] For example, the full-JIF is at least as informative as the all_but_$\boldsymbol{Y}$-JIF for every $\boldsymbol{Y}$, which is at least as informative as the zero-JIF. As one may naturally expect, we can show that, if a less informed individual has an incentive to manipulate an aggregation rule, then she also has an incentive to manipulate the same rule when holding more information. Interestingly though, these incentives do not necessarily coincide. That is, the profile that triggers the manipulation under less information may be different than the one causing the manipulation under more information [consult Reijngoud (2011) for the voting-counterpart of this result].

Proposition 3 *If a JIF π is at least as informative as another JIF σ, then all aggregation rules that are σ-manipulable for a class of preferences PR are also π-manipulable for PR.*

Corollary 4 *If a JIF π is at least as informative as another JIF σ, then all aggregation rules that are π-strategyproof for a class of preferences PR are also σ-strategyproof for PR.*

Hence, we have now settled that for any JIF π, all aggregation rules that are full-strategyproof are also π-strategyproof. Roughly speaking, withholding information from the individuals can never damage the strategyproofness of an aggregation rule. By Theorem 1, this implies that for any JIF π, all independent and monotonic aggregation rules are π-strategyproof for the class of closeness-respecting preferences.

Corollary 5 *An aggregation rule F is π-strategyproof for all JIFs π for the class of closeness-respecting preferences if and only if F is both independent and monotonic.*

Specifically, since quota rules are independent and monotonic (Dietrich and List 2007b), they are π-strategyproof for all JIFs π, for the class of closeness-respecting preferences.

[14] Equivalently, π is at least as informative as σ if for all individuals i and judgment sets J_i the partition on $\mathcal{J}(\Phi)^{n-1}$ induced by π is finer than the one induced by σ.

4.2 Partial information makes a difference

Next, we ponder: Is it possible to get qualitatively different results regarding the manipulability of aggregation rules by introducing the model of partial information? As we will shortly see, the answer is clearly positive. To begin with, we can guarantee the existence of a strategyproof aggregation rule for scenarios where the individuals are missing a reasonable part of the information concerning the judgments of their peers, even when the same rule is manipulable for fully informed individuals.

Theorem 6 *For any agenda Φ, there exist an aggregation rule F and a JIF π different than the zero-JIF, such that F is π-strategyproof but not full-strategyproof for the class of closeness-respecting preferences.*

Theorem 6 ensures that for any agenda Φ, if we have control over the information that a group of individuals has access to, then we can suggest the use of an aggregation rule that no-one will be able to manipulate. Moreover, it is worth stressing that our result does not require totally ignorant individuals; rather, having only two individuals that lack information about the judgments of each other already is a sufficient condition (consult the proof in the "Appendix" for the technical details).

To illustrate, suppose that the members of a political party in Greece need to decide about whether a new secretary should be hired in their central office. Among the party members, everyone would of course like to hire their cousin in case the secretary position is open, but only the party's top members, the leader and the sub-leader, have a realistic chance of doing so. However, these two individuals have a known history of not getting along well (especially when their personal interests are at stake). So, as the decision-making consultant of the party, we could wisely suggest the use of the following aggregation rule for the vote: the secretary position will be approved if and only if exactly one of the two people on the top votes in favour of it (so no fight between the leader and the sub-leader will create instability in the party in case they both want the opening) and at least half of the other members also believe it is a good idea to have the new position (suggesting that it is a reasonable choice to make). Now, if we could also guarantee that the two top members will not have access to each other's judgments about the position before the vote, then we can make sure that no-one who wants the opening to be approved will have an incentive to lie; for instance, if the leader attempts to do so, she would immediately risk to have the position rejected in case the sub-leader does not have a cousin available for hiring at the moment and thus is opposed to the opening. On the other hand, if the leader is informed that the sub-leader is planning to vote positively, then she could untruthfully vote negatively and, after the position is approved, get in the fight about hiring her preferred person.

As a next step, we show that partial information can facilitate strategyproofness also in case there are specific restrictions regarding the choice of the aggregation rule. Remarkably, even if additional constraints (formulated as desirable axiomatic properties) have to be fulfilled by the aggregation rule, immunity to manipulation can nevertheless be achieved. In particular, Theorem 2, the impossibility theorem of Dietrich and List (2007c), breaks down.

Theorem 7 *For any agenda Φ and any number of individuals $n \geq 7$, the plurality rule $F^{p\ell}$ along with any lexicographic tie-breaking rule is nondictatorial, complete, consistent, responsive, and immune to zero-manipulation for the class of closeness-respecting preferences.*

Corresponding results have been proved for the plurality rule in voting, albeit under stronger assumptions regarding the number of individuals (Conitzer et al. 2011; Reijngoud and Endriss 2012). The main insight in the proof of Theorem 7 is that when the plurality rule is applied, an individual has an incentive to manipulate if and only if she knows with certainty that her judgment is pivotal to the achievement of a strictly preferable outcome. When the individual is fully aware of the judgments of her peers, there are profiles where such an incentive is obvious. However, by restraining the information that the individual holds, her incentives to manipulate disappear.

But it is important to emphasise that Theorem 7 does not trivially rely on the features of zero information. To convince the reader of this fact, we inspect the average-voter rule, which satisfies all the demands of Theorem 2 besides strategyproofness, and we show that zero-information does *not* solve the problem of manipulability in this case.

Proposition 8 *There exist an agenda Φ and a group N for which the average-voter rule F^{av} together with a lexicographic tie-breaking order is susceptible to zero-manipulation for the class of closeness-respecting preferences.*

Proposition 8 together with the fact that adding more information can only harm strategyproofness (Proposition 3) makes it evident that there is no class of JIFs for which immunity to manipulation is guaranteed for every rule.

Corollary 9 *There is no JIF π for which every aggregation rule F is π-strategyproof for the class of closeness-respecting preferences.*

5 The premise-based rule

This section examines an aggregation rule that is popular due to its practical use and the properties of which have been extensively studied in the literature on judgment aggregation, namely the premise-based rule (Pettit 2001; Bovens and Rabinowicz 2006; Dietrich and List 2007c; Dietrich and Mongin 2010; Hartmann and Sprenger 2012). We restrict attention to the most relevant agendas for the application of the premise-based rule, namely the conjunctive agendas. However, all our results hold for disjunctive agendas too. Throughout our analysis, we are going to keep assuming that the preferences of the individuals are always complete.

As has become evident in the introduction, the premise-based rule can be manipulated by an individual who wants the conclusion of the conjunctive agenda to be rejected, but who nevertheless truthfully accepts one premise, say premise p. In case such an individual knows that her judgment on p is pivotal concerning the collective decision on the conclusion, i.e., that by untruthfully rejecting p the—previously accepted—conclusion will be rejected by the group, the individual will prefer to lie. This observation was formalised by Dietrich and List (2007c), who showed

that the premise-based rule is full-manipulable for the class of closeness-respecting preferences. In particular, the result of Dietrich and List hinges on individuals with preferences that are only interested in the conclusion of the agenda and completely ignore the collective decision on the premises. Dietrich and List refer to these preferences as *outcome-oriented* (we will call them *conclusion-oriented* instead) and justify them by assuming that only the conclusion and not the premises carries consequences that the individuals care about. Concretely, for every judgment set J_i, let $O(J_i)$ be the set of preferences $\succsim_i$ such that for all judgment sets $J, J' \in 2^{\Phi}$, $J \succsim_i J'$ if and only if J' agreeing on c with J_i implies that J agrees on c with J_i. Then, $O = \{O(J_i) : J_i \in \mathcal{J}(\Phi)\}$ is the class of conclusion-oriented preferences.

Proposition 10 (Dietrich and List 2007c) *For a conjunctive agenda, the premise-based rule is full-manipulable for the class of conclusion-oriented preferences.*

Even though the news concerning the manipulability of the premise-based rule is negative at first sight, several assumptions associated with it deserve further investigation. These are the questions we shall focus on:

- Is the premise-based rule manipulable by individuals who have limited information about the judgments of their peers?
- Do individuals with specific and reasonable types of preferences, different from the conclusion-oriented ones, still have incentives to manipulate the premise-based rule?

Unfortunately, if we address the first question while restricting attention only to individuals with conclusion-oriented preferences, then we continue to obtain a negative result—even under the assumption that the manipulator does not have access to any kind of information.

Lemma 11 *For any conjunctive agenda, the premise-based rule is zero-manipulable for the class of conclusion-oriented preferences.*

By Proposition 3 and the above lemma, we obtain the following generalisation of Proposition 10:

Proposition 12 *For any conjunctive agenda and for any JIF π, the premise-based rule is π-manipulable for the class of conclusion-oriented preferences.*

So, conclusion-oriented individuals have incentives to manipulate under any kind of information that they may hold about the judgments of their peers. This provides us with an extra motivation to study the strategic behaviour of individuals with different preferences, still reasonable for the scenarios addressed by the premise-based rule. Dietrich and List (2007c) initiated this discussion by considering individuals with *reason-based* preferences, i.e., preferences that aim at maximising the agreement between the individual's truthful judgment and the collective decision on the premises only, disregarding the conclusion. Although the work of Dietrich and List (2007c) brought to light a positive result regarding the strategyproofness of the premise-based rule, their assumption of reason-based preferences is rather restrictive. In the sequel, instead we are going to demonstrate different ways of obtaining strong positive results, by exploiting our framework of partial information in combination with various kinds of preferences.

5.1 Conclusion-prioritising preferences

The class of *conclusion-prioritising* preferences P refines the class of conclusion-oriented preferences. Consider a conjunctive agenda Φ, where Φ^p is the set of its premises. The preference relations in P capture the idea that the individuals give highest priority to the outcome on the conclusion, and secondarily, they try to maximise the agreement on the premises. Formally, for each judgment set J_i, let $P(J_i)$ be the set of complete preferences $\succsim_i$ such that for all judgment sets $J, J' \in 2^{\Phi}$, $J \succ_i J'$ if and only if (i) J agrees with J_i on c but J' disagrees with J_i on c , or (ii) both J and J' agree or disagree with J_i on c and $|\Phi^p \cap J \cap J_i| > |\Phi^p \cap J' \cap J_i|$. Then, $P = \{P(J_i) : J_i \in \mathcal{J}(\Phi)\}$ is the class of conclusion-prioritising preferences.

We can show that the premise-based rule is full-manipulable for conclusion-prioritising preferences, similarly to conclusion-oriented preferences. However, under partial information the balance changes. The premise-based rule is immune to manipulation for conclusion-prioritising preferences, while it is still manipulable for conclusion-oriented preferences. Furthermore, the amount of information that needs to be absent in order to achieve strategyproofness is remarkably small. Speaking informally, for large agendas truthfulness is guaranteed even when the individuals know almost everything about the judgments of the rest of the group, i.e., when their uncertainty tends to 0. Before stating our result formally, we define a measure of the *uncertainty* related to a JIF π.

Definition 7 The uncertainty of individual $i \in N$ induced by JIF π in profile $\boldsymbol{J} \in \mathcal{J}(\Phi)^n$ is defined as follows:

$$U_i^{\boldsymbol{J}}(\pi) = \frac{|\mathcal{W}_i^{\pi,\boldsymbol{J}}| - 1}{|\mathcal{J}(\Phi)^{n-1}| - 1} \quad \text{if } |\mathcal{J}(\Phi)| > 1 \quad \text{and} \quad U_i^{\boldsymbol{J}}(\pi) = 0 \text{ otherwise.}$$

The **uncertainty** of a JIF π is the maximal uncertainty it can induce:

$$U(\pi) = \max_{i \in N} \max_{\boldsymbol{J} \in \mathcal{J}(\Phi)^n} U_i^{\boldsymbol{J}}(\pi).$$

Thus, the uncertainty that the JIF π induces for an individual on a profile is a real number between 0 and 1, i.e., $0 \leq U_i^{\boldsymbol{J}}(\pi) \leq 1$, where 0 denotes full certainty and 1 total uncertainty. The more partial profiles are possible for the individuals, the more uncertainty increases. For example, according to the full-JIF the individuals only consider possible the truthful partial profile, thus the uncertainty of the full-JIF is 0. At the other extreme, the uncertainty of the zero-JIF is 1, because according to it the individuals deem possible all the partial profiles.

Lemma 13 *Consider a conjunctive agenda Φ with at least two premises $p_1, p_2 \in \Phi^p$. For $Y = \{p_1, p_2, \neg p_1, \neg p_2\}$ and $\boldsymbol{Y} = (Y, \ldots, Y)$, the premise-based rule is immune to all_but_$\boldsymbol{Y}$-manipulation for the class of conclusion-prioritising preferences.*

Theorem 14 *Consider a conjunctive agenda Φ with at least two premises. The premise-based rule F^{pr} is susceptible to full-manipulation for the class of conclusion-prioritising preferences P. However, there is a family of JIFs $\{\pi^x : x \in \mathbb{N}\}$ with*

$\lim_{x\to\infty} U(\pi^x) = 0$ *such that* F^{pr} *is immune to* π_m*-manipulation for* P*, where* $m = |\Phi|$.

Lemma 13 establishes that, even when the preferences of the individuals prioritise the conclusion in a conjunctive agenda—but when they do not totally overlook the premises—strategyproofness of the premise-based rule is guaranteed in combination with certain modest assumptions on the uncertainty of individuals. This result is important for at least two reasons. First, it shows that caring about the conclusion is not necessarily detrimental to the strategyproofness of the premise-based rule, as the result of Dietrich and List (2007c) seems to imply; and second, it confirms that aiming for zero information is not the only way of achieving positive results. Then, Theorem 14 stresses an additional intriguing observation: that determining exactly how much information causes the manipulability of an aggregation rule can be quite an intricate challenge. Specifically, it proves that for big agendas, a rule can be susceptible to manipulation under full information but strategyproof under almost-full information.

5.2 Hamming-distance preferences

One particular example of commonly used closeness-respecting preferences in the literature are the *Hamming-distance preferences* (Dietrich and List 2007c; Endriss et al. 2012; Baumeister et al. 2015; Botan et al. 2016). These preferences are widely adopted in settings where the individuals are expected to care equally about all the formulas in the agenda. For every individual i, the Hamming-distance naturally induces a (reflexive, transitive and complete) preference relation $\succsim_i$ on judgment sets. Let $H(J_i)$ be the set of preferences $\succsim_i$ such that, for all judgment sets $J, J' \in 2^{\Phi}$ it is the case that $J \succsim_i J'$ if and only if $H(J, J_i) \leq H(J', J_i)$. Then, $H = \{H(J_i) : J_i \in \mathcal{J}(\Phi)\}$ is the class of Hamming-distance preferences.

We now show that, if the individuals have Hamming-distance preferences, then the strategyproofness of the premise-based rule is guaranteed, independently of the amount of information that the individuals possess.

Lemma 15 *For any conjunctive agenda* Φ*, the premise-based rule is immune to full-manipulation for the class of Hamming-distance preferences.*

By Corollary 4 and the above lemma, we thus obtain:

Theorem 16 *For any conjunctive agenda* Φ *and any JIF* π*, the premise-based rule is immune to* π*-manipulation for the class of Hamming-distance preferences.*

Overall, we have underlined the gravity of the assumptions that one makes about the preferences of the individuals in an aggregation scenario as far as the manipulability of aggregation rules is concerned, especially when partial information comes into play. It now has become explicit that the manipulability of the premise-based rule is based on a special subset of the closeness-respecting preferences, namely the conclusion-oriented preferences, and narrowing down our analysis to individuals that completely overlook the conclusion in the agenda is not the only way to achieve strategyproofness.

Notably, the non-independent premise-based rule is strategyproof under full information for a different subclass of the closeness-respecting preferences, consisting of the Hamming-distance preferences (Theorem 16).[15] Moreover, by adding an extra layer of uncertainty, the premise-based rule is also strategyproof for a third subclass of the closeness-respecting preferences, namely the class of the conclusion-prioritising preferences (Theorem 14).

6 Conclusion

We have introduced a novel model of strategic manipulation in judgment aggregation that weakens the standard assumption of every potential manipulator having full information regarding the judgments of her peers. Our technical results clarify the relationship between the standard model of full information and our model of partial information and demonstrate that, by moving to more realistic assumptions regarding the information available to individuals, we can avoid some of the negative results proved in the literature and instead obtain strategyproof judgment aggregation rules. This is true, in particular, for the important premise-based rule, which turns out to be strategyproof under two sets of assumptions that may be deemed reasonable in certain domains.

While we have introduced a general framework for modelling the information available to an individual who might engage in strategic manipulation, our results mostly relate to very specific instances of this general model, such as the case of zero information. In the future, more work will be needed to identify and analyse more realistic choices of judgment information functions that are relevant to specific applications. For example, as previously mentioned, building on the ideas of Chopra et al. (2004) one could assume that individuals are part of a social network and have full information on their immediate neighbours, but no information on those individuals further removed in the network. Finally, an other important direction of research expanding on our work concerns obtaining deeper technical results, for instance characterising for what—if any—types of partial information the impossibility of Dietrich and List (2007c) is reproduced.

Appendix: Proofs

Proof of Theorem 1 We only show the non-trivial direction, which states that if an aggregation rule F is strategyproof for all reflexive, transitive and complete closeness-respecting preferences, then F is independent and monotonic.

[15] This observation relates to a more general discussion by Botan et al. (2016), who remark that (one direction of) the characterisation of Dietrich and List (2007c) fails for certain subsets of closeness-respecting preferences.

We need an intermediate result by Dietrich and List (2007c). These authors define a preference-free notion of strategyproofness. They call the aggregation rule F *(preference-free) manipulable* at the profile of judgments $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$ by individual i if there is a formula φ in the pre-agenda Φ^+ such that $F(\boldsymbol{J})$ disagrees with J_i on φ, but $F(J_i^*, \boldsymbol{J}_{-i})$ agrees with J_i on φ, for some untruthful judgment J_i^*. Based on this definition, Dietrich and List (2007c) prove that every aggregation rule is immune to (preference-free) manipulation if and only if it is independent and monotonic.

Back to our proof, it now suffices to demonstrate that if an aggregation rule F is strategyproof for all reflexive, transitive and complete closeness-respecting preferences, then F is immune to (preference-free) manipulation. So, assume that strategyproofness is the case. To show immunity to (preference-free) manipulation, consider a formula $\varphi \in \Phi$, an individual $i \in N$, and a truthful profile $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$ such that $F(J_i, \boldsymbol{J}_{-i})$ disagrees with J_i on φ. We need to prove that $F(J_i^*, \boldsymbol{J}_{-i})$ still disagrees with J_i on φ, for every dishonest judgment J_i^*. We define a preference relation $\succsim_i$ over all possible collective outcomes such that $J \succsim_i J'$ if and only if J_i agrees on φ with J but not with J', or J_i agrees on φ with both J and J', or it disagrees with both. Intuitively, this would be the case if individual i only cares about the formula φ in the agenda. It is easy to verify that $\succsim_i$ is reflexive, transitive, complete, and closeness-respecting. Hence, by strategyproofness it will be $F(J_i, \boldsymbol{J}_{-i}) \succsim_i F(J_i^*, \boldsymbol{J}_{-i})$ for all untruthful judgments J_i^*. But as $F(J_i, \boldsymbol{J}_{-i})$ disagrees with J_i on φ, the definition of $\succsim_i$ implies that $F(J_i^*, \boldsymbol{J}_{-i})$ also disagrees with J_i on φ, for every dishonest judgment J_i^*. □

Proof of Proposition 3 Consider a JIF π that is at least as informative as another JIF σ and an aggregation rule F that is σ-manipulable. We will show that F is also π-manipulable. By σ-manipulability we know that there are an individual $i \in N$, a profile $\boldsymbol{J} = (J_1, \ldots, J_n) \in \mathcal{J}(\Phi)^n$, a preference relation $\succ_i \in PR(J_i)$, and an insincere judgment J_i^* such that (i) $F(J_i^*, \boldsymbol{J'}_{-i}) \succ_i F(J_i, \boldsymbol{J'}_{-i})$, for some $\boldsymbol{J'}_{-i} \in \mathcal{W}_i^{\sigma,\boldsymbol{J}}$, and (ii) $F(J_i^*, \boldsymbol{J''}_{-i}) \succsim_i F(J_i, \boldsymbol{J''}_{-i})$, for all other $\boldsymbol{J''}_{-i} \in \mathcal{W}_i^{\sigma,\boldsymbol{J}}$.

Now, consider the profile $\boldsymbol{J'} = (J_i, \boldsymbol{J'}_{-i})$. We will show that if the truthful profile is $\boldsymbol{J'}$, then the individual i has an incentive to π-manipulate by reporting the untruthful judgment set J_i^*. It suffices to show two things: (i') $\boldsymbol{J'}_{-i} \in \mathcal{W}_i^{\pi,\boldsymbol{J'}}$ and (ii') $\mathcal{W}_i^{\pi,\boldsymbol{J'}} \subseteq \mathcal{W}_i^{\sigma,\boldsymbol{J}}$. Indeed, the former means that i considers $\boldsymbol{J'}_{-i}$ possible, which together with (i) means that there is a possible profile for which she would be strictly better off. And the latter means that (ii) also holds for all $\boldsymbol{J''}_{-i} \in \mathcal{W}_i^{\pi,\boldsymbol{J'}}$.

But we know that (i') holds, due to the reflexivity of $\mathcal{W}$. So it remains to prove (ii'). In fact, since $\mathcal{W}_i^{\pi,\boldsymbol{J'}} \subseteq \mathcal{W}_i^{\sigma,\boldsymbol{J'}}$ due to π being at least informative as σ, it suffices to show $\mathcal{W}_i^{\sigma,\boldsymbol{J'}} \subseteq \mathcal{W}_i^{\sigma,\boldsymbol{J}}$. So take any $\boldsymbol{J''}_{-i} \in \mathcal{W}_i^{\sigma,\boldsymbol{J'}} = \mathcal{W}_i^{\sigma,(J_i,\boldsymbol{J'}_{-i})}$. Together with $\boldsymbol{J'}_{-i} \in \mathcal{W}_i^{\sigma,(J_i,\boldsymbol{J}_{-i})}$, which holds due reflexivity, an application of transitivity yields $\boldsymbol{J''}_{-i} \in \mathcal{W}_i^{\sigma,(J_i,\boldsymbol{J}_{-i})} = \mathcal{W}_i^{\sigma,\boldsymbol{J}}$ and we are done. □

Proof of Theorem 6 Consider a formula $\psi \in \Phi$, two individuals $i, j \in N$, and the aggregation rule F such that for all profiles $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$ and formulas $\varphi \in \Phi$, it is the case that (i) $\varphi \in F(\boldsymbol{J})$ if and only if $|N_\varphi^{\boldsymbol{J}}| \geq \frac{n}{2}$ for all $\varphi \neq \psi$ and (ii) $\psi \in F(\boldsymbol{J})$ if and only if $(\psi \in J_i \backslash J_j$ or $\psi \in J_j \backslash J_i)$ and $|\{k \in N \backslash \{i, j\} \mid \psi \in J_k\}| \geq \frac{n-2}{2}$. Loosely speaking, F requires the majority's acceptance for all "standard" formulas,

and regarding the "special" formula ψ, F can only accept it when exactly one of the "special" individuals i, j do so, and in addition the majority of the rest of the individuals approve it.[16] Now, take the JIF π under which each individual $k \in N\backslash\{i, j\}$ is completely informed about the judgments of the others, while individuals i and j are only ignorant about each other's opinion (and fully know all other judgments). Formally, π is the besides_$\boldsymbol{I}$-JIF, where $\boldsymbol{I} = (I_1, \ldots, I_n)$ is such that $I_k = \emptyset$ for all $k \in N\backslash\{i, j\}$, $I_i = \{j\}$ and $I_j = \{i\}$. The aggregation rule F is obviously not monotonic, so we know by Theorem 1 that it is not full-strategyproof for the class of closeness-respecting preferences. However, we can show that F is π-strategyproof for the same class of preferences.

Clearly, the individuals in N besides i and j have no incentives to manipulate (since the rule is independent and also behaves monotonically with respect to their judgments). Consider now, without loss of generality, individual i, a profile $(J_i, \boldsymbol{J}_{-i})$, a judgment set J_i^*, and a preference relation $\succsim_i \in C(J_i)$. Furthermore, suppose that there is a judgment set J_i^* such that $F(J_i^*, \boldsymbol{J}'_{-i}) \succ_i F(J_i, \boldsymbol{J}'_{-i})$, for some $\boldsymbol{J}'_{-i} \in \mathcal{W}_i^{\pi,\boldsymbol{J}}$. By the definition of closeness-respecting preferences, this means that there is some $\varphi \in J_i$ such that $\varphi \in F(J_i^*, \boldsymbol{J}'_{-i})$ and $\varphi \notin F(J_i, \boldsymbol{J}'_{-i})$. But by the definition of the rule F this can only happen if $\varphi = \psi$, $\psi \notin J_i^*$, and $\psi \in J'_j$ (where J'_j is the judgment set of individual j in profile $\boldsymbol{J}'_{-i}$). Then, since $\psi \notin J_i^*$, we know that $\psi \neq \top$. So there exists a model M of propositional logic such that $M \not\models \psi$. Hence we can consider a different (complete and consistent) judgment set of individual j, which individual i deems possible, based on the formulas that are verified by M: $J''_j = \{\varphi \in \Phi : M \models \varphi\}$. Then, $\psi \notin J''_j$. Preserving the judgments of the rest of the group in the partial profile $\boldsymbol{J}'_{-i}$ and replacing the judgment of individual j, with J''_j, we have the new partial profile $\boldsymbol{J}''_{-i} \in \mathcal{W}_i^{\pi,\boldsymbol{J}}$. Thus by the definition of the rule F, it holds that $F(J_i, \boldsymbol{J}''_{-i}) = F(J_i^*, \boldsymbol{J}'_{-i})$ and $F(J_i^*, \boldsymbol{J}''_{-i}) = F(J_i, \boldsymbol{J}''_{-i})$, which means that $F(J_i, \boldsymbol{J}''_{-i}) \succ_i F(J_i^*, \boldsymbol{J}''_{-i})$. To sum up, we found a possible judgment of individual j that will make the risk-averse individual i unwilling to manipulate. Using the same argument, we can prove that individual j does not have an incentive to manipulate either, and we conclude that our rule is strategyproof. □

Proof of Theorem 7 Suppose that the number of individuals n is odd, $n = 2k + 1$, for some integer $k \geq 3$ (the case for even n is analogous). The properties of nondictatorship, completeness, consistency and responsiveness are easily checked to be satisfied for the plurality rule $F^{p\ell}$ together with a lexicographic tie-breaking rule. Thus, we only have to show that $F^{p\ell}$ is immune to zero-manipulation for the class of closeness-respecting preferences. Consider an arbitrary individual i, a profile $(J_i, \boldsymbol{J}_{-i})$, and a closeness-respecting preference $\succsim_i \in C(J_i)$, and suppose that there is a judgment set J_i^* such that $F^{p\ell}(J_i^*, \boldsymbol{J}'_{-i}) \succ_i F^{p\ell}(J_i, \boldsymbol{J}'_{-i})$, for some partial profile $\boldsymbol{J}'_{-i}$ (otherwise the rule is already immune to manipulation). By definition of closeness-respecting preferences and the plurality rule, this can happen only if the collective outcome $F^{p\ell}(J_i, \boldsymbol{J}'_{-i})$ induced by individual i's truthful judgment is some judgment set J and the manipulated result $F^{p\ell}(J_i^*, \boldsymbol{J}'_{-i})$ is the judgment set J_i^*. Thus, it must be the case

[16] For a motivating scenario where this would be an applicable rule, see the discussion following Theorem 6 in the body of the paper.

that $J_i^* \succ_i J$. Moreover, due to $\succsim_i$ being closeness-respective, we have that $J_i \succsim_i J_i^*$, and by transitivity of preferences it holds that $J_i \succ_i J$. We distinguish the following two cases, to account for all tie-breaking rules.

Case 1 The tie-breaking rule ranks J_i above J. Consider the profile $\boldsymbol{J}'' = (J_i, \boldsymbol{J}''_{-i})$, where k individuals (including i) submit the judgment set J_i, one reports J_i^*, and k other individuals submit the judgment set J. Then, $F^{p\ell}(J_i, \boldsymbol{J}''_{-i}) = J_i$. However, if individual i were to report the insincere judgment J_i^*, there would be $k-1$ submissions of J_i, two of J_i^*, and k submissions of J, leading to $F^{p\ell}(J_i^*, \boldsymbol{J}''_{-i}) = J$. We conclude that $F^{p\ell}(J_i, \boldsymbol{J}''_{-i}) \succ_i F^{p\ell}(J_i^*, \boldsymbol{J}''_{-i})$, so i will not be willing to manipulate by reporting the untruthful judgment J_i^*.[17]

Case 2 The tie-breaking rule ranks J above J_i. Then, consider the profile where $k+1$ individuals submit the judgment set J_i, no-one submits J_i^*, and k individuals submit J. The proof proceeds as in case 1. □

Proof of Proposition 8 The idea goes as follows. Imagine that some individual in the group has only one undesirable judgment set and this judgment set evaluates all the formulas in the exact opposite way from some dishonest judgment J of hers. If this individual decides to untruthfully submit J without knowing the judgments of her peers, then she can be sure that the Hamming distance between the collective outcome and her unwanted outcome can never be strictly smaller than when she tells the truth. Thus, reporting J can never harm the individual when the average-voter rule is applied, while it can still make her better off in some possible scenario.

Such a case can be materialised if we consider a group of three individuals[18] and an agenda $\Phi = \{\varphi_1, \varphi_2, \varphi_3, \neg\varphi_1, \neg\varphi_2, \neg\varphi_3\}$ such that $\mathcal{J}(\Phi) = \{\{\varphi_1, \varphi_2, \neg\varphi_3\}, \{\varphi_1, \neg\varphi_2, \neg\varphi_3\}, \{\varphi_1, \neg\varphi_2, \varphi_3\}, \{\neg\varphi_1, \varphi_2, \varphi_3\}\}$.[19] Moreover, let us consider the tie-breaking rule based on the linear order $\{\neg\varphi_1, \varphi_2, \varphi_3\} > \{\varphi_1, \varphi_2, \neg\varphi_3\} > \{\varphi_1, \neg\varphi_2, \varphi_3\} > \{\varphi_1, \neg\varphi_2, \neg\varphi_3\}$. Then, assume that individual 1 holds the truthful judgment $J_1 = \{\varphi_1, \varphi_2, \neg\varphi_3\}$ and the closeness-respecting preference $\succsim_1$ such that $\{\varphi_1, \varphi_2, \neg\varphi_3\} \sim_1 \{\varphi_1, \neg\varphi_2, \neg\varphi_3\} \sim_1 \{\varphi_1, \neg\varphi_2, \varphi_3\} \succ_1 \{\neg\varphi_1, \varphi_2, \varphi_3\}$. Now, suppose that the profile $\boldsymbol{J} = (J_1, J_2, J_3)$ is such that $J_2 = \{\neg\varphi_1, \varphi_2, \varphi_3\}$ and $J_3 = \{\varphi_1, \neg\varphi_2, \varphi_3\}$. Then, $F^{av}(\boldsymbol{J}) = \{\neg\varphi_1, \varphi_2, \varphi_3\}$. However, for the profile $\boldsymbol{J}^* = (J_1^*, J_2, J_3)$ where $J_1^* = \{\varphi_1, \neg\varphi_2, \neg\varphi_3\}$, it holds that $F^{av}(\boldsymbol{J}^*) = \{\varphi_1, \neg\varphi_2, \varphi_3\} \succ_i F^{av}(\boldsymbol{J})$. Finally, just by doing the calculations, one can be persuaded that $F^{av}(J_1^*, J_2', J_3') \succsim_1 F^{av}(J_1, J_2', J_3')$, for all other partial profiles $(J_2', J_3') \in \mathcal{J}(\Phi)^2$. □

Proof of Lemma 11 Consider an individual i that truthfully rejects the conclusion of the conjunctive agenda, but accepts some premise p in it. Suppose further that i is

[17] Note that for this part of the proof to work, the assumption that $k \geq 3$ (which means that $n \geq 7$) is necessary. Suppose that $k = 2$, the closeness-respecting preference $\succsim_i$ is such that $J_i \sim_i J_i^*$, and the tie-breaking rule prioritises J_i^* over J. Then, in the profile $\boldsymbol{J}'' = (J_i, \boldsymbol{J}''_{-i})$ two individuals would submit J_i, one would submit J_i^*, and two other individuals would report J, so $F^{p\ell}(J_i, \boldsymbol{J}''_{-i}) = J_i$. Then, in case individual i reported J_i^* instead of J_i, it would hold that $F^{p\ell}(J_i^*, \boldsymbol{J}''_{-i}) = J_i^* \sim_i F^{p\ell}(J_i, \boldsymbol{J}''_{-i})$.

[18] The same proof would work for any number of individuals that is a multiple of three.

[19] We know that such an agenda can be constructed by Dokow and Holzman (2009).

conclusion-oriented, which means that she desires a collective rejection of the conclusion. If individual i has zero information about the judgments of her peers, that is, if she deems all profiles possible to be submitted by the group, then she has an incentive to reject all the premises. Indeed, if the conclusion was already rejected by the truthful profile of the group, individual i has nothing to lose—with less support on some of the premises, the conclusion will still be rejected. But there is always a possible profile for which the conclusion is accepted in case the individual is honest, while it is rejected if the individual lies. To see this, assume that the number of individuals is odd: $n = 2k + 1$ (the proof for an even number of individuals is analogous). Then, consider the profile where all of i's peers accept all the other premises besides p, and exactly k of them reject p. Since $k + 1$ individuals (including i) accept p, all the premises will be accepted by the group and hence the conclusion will be accepted too. However, if i were to reject p, then a majority of individuals would reject p and the conclusion would be rejected as well. □

Proof of Lemma 13 Consider an arbitrary individual i. If i truthfully accepts the conclusion in the agenda, then obviously she has no incentive to manipulate. Suppose, then, that i truthfully rejects the conclusion. If i truthfully rejects all the premises, then she has no incentive to manipulate either. So, we assume without loss of generality that i accepts some premise p, while she rejects the conclusion. Then, the only way that i could become better off is by altering a collective decision that accepts the conclusion, by untruthfully rejecting p and thus making the group reject p too. Now consider $p' \in \{p_1, p_2\} \backslash \{p\}$. Since i does not know the judgments of her peers on p', she considers possible the case where everyone else rejects p'. In such a scenario, the collective outcome would already agree with her on (i.e., reject) the conclusion. So in this case, if individual i rejects p, she will not affect the collective outcome regarding the conclusion. However, she will cause a decrease on the number of premises that the group's judgment and her own individual judgment agree on. This means that individual i has no incentive to manipulate. □

Proof of Theorem 14 The premise-based rule F^{pr} is susceptible to manipulation for the class of conclusion-prioritising preferences P under full information, because there is a profile where an individual i can change the collective result on the conclusion from disagreeing with her truthful judgment to agreeing with it by lying on a premise (see the proof of Lemma 11 for the construction of such a profile). However, as we will see next, F^{pr} is immune to manipulation under partial information.

We construct a family of JIFs $\{\pi^x : x \in \mathbb{N}\}$, where each π^x is defined based on an agenda X with size x as follows. Take an arbitrary conjunctive agenda X with size x and at least two premises. Fixing two arbitrary premises $p_1, p_2 \in X$, we define $Y = \{p_1, p_2, \neg p_1, \neg p_2\}$, and π^x equal to the all_but_$\boldsymbol{Y}$-JIF. Then, by Lemma 13 we have that the premise-based rule is immune to π^x-manipulation. Moreover, we will show that $\lim_{x\to\infty} U(\pi^x) = 0$. We observe that when x tends to infinity, the number of all the possible (partial) profiles on an agenda X with size x tends to infinity too, i.e., $\lim_{x\to\infty} |\mathcal{J}(X)^{n-1}| = \infty$. Let us now consider an individual i and a profile $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$. The number of all the partial profiles that individual i deems possible according to π^x is finite. Specifically, $|\mathcal{W}_i^{\pi^x, \boldsymbol{J}}| = 2^{2(n-1)}$, since only the judgments

of the other $n-1$ individuals with regard to the premises p_1 and p_2 are unknown to individual i. Thus,

$$\lim_{x\to\infty} U_i^{\boldsymbol{J}}(\pi^x) = \lim_{x\to\infty} \frac{\left|\mathcal{W}_i^{\pi^x,\boldsymbol{J}'}\right| - 1}{|\mathcal{J}(X)^{n-1}| - 1} = \lim_{x\to\infty} \frac{4^{n-1}-1}{|\mathcal{J}(X)^{n-1}| - 1} = 0.$$

Finally, since i and $\boldsymbol{J}$ were arbitrary, we have that $\lim_{x\to\infty} U(\pi^x) = 0$. □

Proof of Lemma 15 Suppose, aiming for a contradiction, that there exist an individual i with Hamming-distance preferences $\succsim_i$ and a profile $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$ in which i has an incentive to manipulate. Then there is a judgment set J_i^* such that $F^{pr}(J_i^*, \boldsymbol{J}_{-i}) \succ_i F^{pr}(J_i, \succ J_{-i})$, which by definition of the Hamming-distance preferences means that the judgment set $F^{pr}(J_i^*, \boldsymbol{J}_{-i})$ has strictly more formulas in common with J_i than the judgment set $F^{pr}(J_i, \boldsymbol{J}_{-i})$ has. But according to the premise-based rule, if individual i switches from reporting her truthful judgment J_i to reporting the untruthful judgment J_i^*, it is not possible to obtain a collective decision that is agreeing on a premise with J_i if the initial collective judgment was not agreeing on that premise with J_i too. Hence, the only way for $F^{pr}(J_i^*, \boldsymbol{J}_{-i})$ to have a formula in common with J_i that $F^{pr}(J_i, \boldsymbol{J}_{-i})$ does not have is if that formula is the conclusion. However, in order to achieve this, J_i^* should be untruthful and change the collective judgment on at least one of the premises that J_i and $F^{pr}(J_i, \boldsymbol{J}_{-i})$ agree on. In total, $F^{pr}(J_i^*, \boldsymbol{J}_{-i})$ cannot have strictly more formulas in common with J_i than the judgment set $F^{pr}(J_i, \boldsymbol{J}_{-i})$ has, thereby contradicting our initial assumption. □

References

Baumeister D, Erdélyi G, Erdélyi OJ, Rothe J (2015) Complexity of manipulation and bribery in judgment aggregation for uniform premise-based quota rules. Math Soc Sci 76:19–30

Baumeister D, Rothe J, Selker A-K (2017) Strategic behavior in judgment aggregation. In: Endriss U (ed) Trends in Computational Social Choice, chapter 8. AI Access, pp 145–168

Botan S, Novaro A, Endriss U (2016) Group manipulation in judgment aggregation. In: Proceedings of the 15th international conference on autonomous agents and multiagent systems (AAMAS), pp 411–419

Bovens L, Rabinowicz W (2006) Democratic answers to complex questions–an epistemic perspective. Synthese 150(1):131–153

Bozbay I, Dietrich F, Peters H (2014) Judgment aggregation in search for the truth. Games Econ Behav 87:571–590

Caragiannis I, Procaccia AD, Shah N (2014) Modal ranking: a uniquely robust voting rule. In: Proceedings of the 28th AAAI conference on artificial intelligence, pp 616–622

Chapman B (2002) Rational aggregation. Politics Philos Econ 1(3):337–354

Chopra S, Pacuit E, Parikh R (2004) Knowledge-theoretic properties of strategic voting. In: Proceedings of the 8th European conference on logics in artificial intelligence (JELIA), pp 18–30

Conitzer V, Walsh T (2016) Barriers to manipulation in voting. In: Brandt F, Conitzer V, Endriss U, Lang J, Procaccia AD (eds) Handbook of computational social choice. Cambridge University Press, Cambridge, pp 127–145

Conitzer V, Walsh T, Xia L (2011) Dominating manipulations in voting with partial information. In: Proceedings of the 25th AAAI conference on artificial intelligence, pp 638–643

de Haan R (2017) Complexity results for manipulation, bribery and control of the Kemeny judgment aggregation procedure. In: Proceedings of the 16th international conference on autonomous agents and multiagent systems (AAMAS), pp 1151–1159

Dietrich F (2007) A generalised model of judgment aggregation. Soc Choice Welf 28(4):529–565

Dietrich F (2016) Judgment aggregation and agenda manipulation. Games Econ Behav 95(C):113–136
Dietrich F, List C (2007a) Arrow's Theorem in judgment aggregation. Soc Choice Welf 29(1):19–33
Dietrich F, List C (2007b) Judgment aggregation by quota rules: majority voting generalized. J Theor Politics 19(4):391–424
Dietrich F, List C (2007c) Strategy-proof judgment aggregation. Econ Philos 23(03):269–300
Dietrich F, List C (2008) Judgment aggregation without full rationality. Soc Choice Welf 31(1):15–39
Dietrich F, Mongin P (2010) The premiss-based approach to judgment aggregation. J Econ Theory 145(2):562–582
van Ditmarsch H, Lang J, Saffidine A (2013) Strategic voting and the logic of knowledge. In: Proceedings of the 14th conference on theoretical aspects of rationality and knowledge (TARK), pp 196–205
Dokow E, Holzman R (2009) Aggregation of binary evaluations for truth-functional agendas. Soc Choice Welf 32(2):221–241
Elster J (1998) Deliberative democracy. Cambridge University Press, Cambridge
Endriss U (2016) Judgment aggregation. In: Brandt F, Conitzer V, Endriss U, Lang J, Procaccia AD (eds) Handbook of computational social choice. Cambridge University Press, Cambridge, pp 399–426
Endriss U, Grandi U (2014) Binary aggregation by selection of the most representative voter. In: Proceedings of the 28th AAAI conference on artificial intelligence, pp 668–674
Endriss U, Grandi U, Porello D (2012) Complexity of judgment aggregation. J Artif Intell Res 45(1):481–514
Endriss U, Obraztsova S, Polukarov M, Rosenschein JS (2016) Strategic voting with incomplete information. In: Proceedings of the 25th international joint conference on artificial intelligence (IJCAI), pp 236–242
Gaertner W (2001) Domain conditions in social choice theory. Cambridge University Press, Cambridge
Gärdenfors P (2006) A representation theorem for voting with logical consequences. Econ Philos 22(2):181–190
Gibbard A (1973) Manipulation of voting schemes: a general result. Econometrica 41(4):587–601
Gilboa I (2009) Theory of decision under uncertainty. Cambridge University Press, Cambridge
Grossi D, Pigozzi G (2014) Judgment aggregation: a primer. synthesis lectures on artificial intelligence and machine learning. Morgan & Claypool Publishers, San Rafael
Hartmann S, Sprenger J (2012) Judgment aggregation and the problem of tracking the truth. Synthese 187(1):209–221
Jeffrey R (1983) Bayesianism with a human face. Test Sci Theor 10:133–156
List C (2003) A possibility theorem on aggregation over multiple interconnected propositions. Math Soc Sci 45(1):1–13
List C (2012) The theory of judgment aggregation: an introductory review. Synthese 187(1):179–207
List C, Pettit P (2002) Aggregating sets of judgments: an impossibility result. Econ Philos 18(1):89–110
List C, Puppe C (2009) Judgment aggregation: a survey. In: Anand P, Pattanaik P, Puppe C (eds) Handbook of rational and social choice. Oxford University Press, Oxford, pp 457–482
Meir R, Lev O, Rosenschein JS (2014) A local-dominance theory of voting equilibria. In: Proceedings of the 15th ACM conference on economics and computation (EC), pp 313–330. ACM, New York
Nehring K, Puppe C (2007) The structure of strategy-proof social choice–part I: general characterization and possibility results on median spaces. J Econ Theory 135(1):269–305
Obraztsova S, Markakis E, Thompson DR (2013) Plurality voting with truth-biased agents. In: Proceedings of the 6th international symposium on algorithmic game theory (SAGT), pp 26–37
Osborne MJ, Rubinstein A (2003) Sampling equilibrium, with an application to strategic voting. Games Econ Behav 45(2):434–441
Pettit P (2001) Deliberative democracy and the discursive dilemma. Philos Issues 11(1):268–299
Ray P (1973) Independence of irrelevant alternatives. Econometrica 41(5):987–991
Reijngoud A (2011) Voter response to iterated poll information. Master's thesis, ILLC, University of Amsterdam
Reijngoud A, Endriss U (2012) Voter response to iterated poll information. In: Proceedings of the 11th international conference on autonomous agents and multiagent systems (AAMAS), pp 635–644
Satterthwaite MA (1975) Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. J Econ Theory 10(2):187–217
Sen A (1969) Quasi-transitivity, rational choice and collective decisions. Rev Econ Stud 36(3):381–393
Sen A (1970) Collective choice and social welfare. Holden-Day, San Francisco
Sen A (1973) Behaviour and the concept of preference. Economica 40(159):241–259

Terzopoulou Z (2017) Manipulating the manipulators: richer models of strategic behavior in judgment aggregation. Master's thesis, ILLC, University of Amsterdam

Terzopoulou Z, Endriss U, de Haan R (2018) Aggregating incomplete judgments: axiomatisations for scoring rules. In: Proceedings of the 7th international workshop on computational social choice (COMSOC)

Zwicker WS (2016) Introduction to the theory of voting. In: Brandt F, Conitzer V, Endriss U, Lang J, Procaccia AD (eds) Handbook of computational social choice. Cambridge University Press, Cambridge, pp 23–56