Reliable Methods of Judgment Aggregation

Stephan Hartmann *, Gabriella Pigozzi †, and Jan Sprenger ‡

May 7, 2009

Abstract

The aggregation of consistent individual judgments on logically interconnected propositions into a collective judgment on the same propositions has recently drawn much attention. Seemingly reasonable aggregation procedures, such as propositionwise majority voting, cannot ensure an equally consistent collective conclusion. The literature on judgment aggregation refers to such a problem as the *discursive dilemma*. In this paper we assume that the decision which the group is trying to reach is factually right or wrong. Hence, we address the question of how good various approaches are at selecting the right conclusion. We focus on two approaches: distance-based procedures and a Bayesian analysis. They correspond to group-internal and group-external decision-making, respectively. We compare those methods in a probabilistic model whose assumptions are subsequently relaxed. The findings vindicate that in judgment aggregation problems, (i) reasons should carry higher weight in the voting procedure than the conclusion, and (ii) considering members of an advisory board to be highly competent is a better strategy than to underestimate their advice.

1 Introduction

Judgment aggregation [7, 8, 11] is an emerging research area in economics. It investigates how to aggregate individual judgments on logically related propositions to a group judgment on those propositions. Examples of groups that need to aggregate individual judgments are expert panels, legal courts, boards, and councils. The propositions are of two kinds: *premises* and a *conclusion*. The first serve as supporting reasons to derive a judgment on the conclusion. Consider, for example, a city council that has to make a decision on whether to build a new harbor site (represented by a proposition C, the conclusion). This project is eligible for public funding if and only if two premises are satisfied: first, there is sufficient request for new harbor sites that cannot be met by existing harbor sites (represented by proposition A_1), and second, the nearby marine reserve is not badly affected (represented by proposition A_2). The decision rule can be

*Tilburg Center for Logic and Philosophy of Science, Tilburg University, P.O. Box 90153, 5000 LE Tilburg. E-mail: s.hartmann@uvt.nl.

†Individual and Collective Reasoning, Computer Science and Communication, University of Luxembourg, L-1359 Luxembourg. E-mail: gabriella.pigozzi@uni.lu.

‡Tilburg Center for Logic and Philosophy of Science, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, E-mail: j.sprenger@uvt.nl.

	A_1	A_2	C
Agents 1, 2, 3	Yes	Yes	Yes
Agents 4, 5	Yes	No	No
Agents 6, 7	No	Yes	No
Majority	Yes	Yes	No

Table 1: An illustration of the discursive dilemma under the aggregation rule $(A_1 \wedge A_2) \leftrightarrow C$.

formally expressed as the formula $(A_1 \wedge A_2) \leftrightarrow C$. Each member of the council expresses her judgment on A_1, A_2 and C such that the rule $(A_1 \wedge A_2) \leftrightarrow C$ is satisfied.

How shall we derive a group judgment given the individuals' opinions on premises and conclusion? It is assumed that each individual expresses judgments on the propositions while respecting the logical relations. If we define the group opinion as the majority view on the issues (premises and conclusion), it turns out that the group may take an inconsistent position, as shown in Table 1. The city council may face a situation where the majority thinks that the new harbor site should not be built. However, it will not be possible to provide *reasons* for this judgment as a majority of the members agrees that there is sufficient request for further harbor sites and another majority agrees that the nearby marine reserve is not badly affected. The literature on judgment aggregation refers to such a problem as the *discursive dilemma*.

Two ways to avoid the paradox are the *premise-based procedure* (PBP) and the *conclusion-based procedure* (CBP) [2, 14]. According to PBP, each member expresses her judgment on each premise. The conclusion is then inferred from the rule $(A_1 \wedge A_2) \leftrightarrow C$ and from the judgment of the group majority on A_1 and A_2. If the individuals of the example followed the PBP, the new harbor site would be approved of. On the other hand, according to the CBP, the members privately decide on A_1 and A_2 and only express their opinions on C publicly. The judgment of the group is then inferred from applying the majority rule to the individual judgments on C. In the above example, the application of the CBP would stop the harbor project, contradicting the results of PBP.

The relevance of such aggregation problems goes beyond the specific city council example, because it applies to all situations where individual binary evaluations need to be combined into a group decision. Furthermore, the problem of aggregating individual judgments is not restricted to majority voting, but it applies to *all* aggregation procedures satisfying some seemingly desirable conditions. For an overview on the impossibility results and on the characterizations of aggregation rules and decision problems provided by the literature, the reader is referred to [12].

In this paper we assume that a group judgment is factually right or wrong and address the question of how *reliable* various approaches are at selecting the right conclusion. In other words, we ask whether some natural and plausible methods can be trusted in practice, generalizing the approach of Bovens and Rabinowicz [1] and of List [9] (see also [10]). We adopt two different perspectives, both common and important in public decision-making: First we compare the performance of a group of methods that can be characterized as *distance-*

based procedures, among them the majority fusion operator [5, 6, 15]. They are essentially functions from a N-tuple of individual judgment sets to a group judgment on a conclusion. Here, the crucial question is how much weight the judgments on the conclusion should get, compared to the judgments on the premises. Second, we ask how an external decision-maker would aggregate the group members' judgments if she took additional information into account, e.g. assessments of the group members' individual competence or the conclusion's (prior) probability. This group-external perspective is implemented by means of a full *Bayesian analysis* and serves as a benchmark for the distance-based procedures. Moreover, we extract recommendations for group-external judgment aggregation under incomplete, or even misleading information.

We analyze the reliability of judgment aggregation procedures in a probabilistic model, relying on analytical results as well as numerical simulations. In particular, we embed CBP, PBP and the majority fusion operator, into a family of *distance-based procedures* (Section 2). We study the properties of that family of aggregation methods and demonstrate the robustness of our results by relaxing the model assumptions (Section 3). Then we review the results of section 3 from a Bayesian perspective and compare group-internal (distance-based) to group-external aggregation procedures (Section 4). Finally, we derive recommendations for policy-making and sum up the main insights (Section 5).

2 The General Model: Distance-Based Procedures

The premise- and the conclusion-based procedures represent two extremes: PBP is only a function of the judgments on the premises, CBP is only a function of the judgments on the conclusion. This prompts the question of whether we can give a general description of aggregation procedures which combine the judgments on premises and conclusion, with PBP and CBP as extreme points.

A typical aggregation procedure which considers all elements of a judgment set is the *majority fusion operator* (MFO). MFO represents all positive verdicts with one, and all negative verdicts with zero (see [15] for a precise definition). Table 2 illustrates MFO in the harbor example from the introduction – the judgment set which comes closest to the group average, calculated by componentwise distances, is selected by MFO. Thus, each premise has the same weight as the conclusion in determining the group outcome.

The idea to use distance-based approaches as a method to resolve inconsistencies stems from computer science [5, 6]: the intuition is that an inconsistent database is replaced by the consistent database that comes "closest" to it. Pigozzi [15] was the first to apply distance-based methods to the problem of judgment aggregation.

We generalize Pigozzi's approach and present a continuum of distance-based procedures, parametrized by a real number $t \geq 0$ which contains PBP and CBP as extremes. The introductory example of two premises is generalized to an aggregation problem where the conclusion is satisfied if and only if M premises are jointly satisfied. In other words, we deal with the aggregation rule $(A_1 \wedge A_2 \wedge \ldots \wedge A_M) \leftrightarrow C$ where the A_i denote the premises and C denotes the conclusion. Notably, our analysis extends to other truth-functional combinations of the

	A_1	A_2	C	Total
Agents 1,2,3	1	1	1	—
Agents 4,5	1	0	0	—
Agents 6,7	0	1	0	—
Average	5/7	5/7	3/7	—
Distance to (1,1,1)	2/7	2/7	4/7	8/7
Distance to (1,0,0)	2/7	5/7	3/7	10/7
Distance to (0,1,0)	5/7	2/7	3/7	10/7
Distance to (0,0,0)	5/7	5/7	3/7	13/7

Table 2: The majority fusion operator in the harbor example of Table 1. It turns out that $\{A_1, A_2, C\}$ is the consistent judgment set that comes closest to the average of the group members' judgments.

premises as well, e.g. the disjunctive aggregation rule $(A_1 \vee A_2 \vee \ldots \vee A_M) \leftrightarrow C$ because this rule is equivalent to $(\neg A_1 \wedge \neg A_2 \wedge \ldots \wedge \neg A_M) \leftrightarrow \neg C$. In general, our analysis captures all aggregation rules where the conclusion is true (false) in only one consistent judgment set.

Therefore, we restrict ourselves in the sequel to the conjunctive aggregation rule $(A_1 \wedge A_2 \wedge \ldots \wedge A_M) \leftrightarrow C$. 2^M judgment sets are consistent with it. We represent them as sets of propositional letters and order them as follows:

$$\begin{aligned} s_0 &= \{A_1, A_2, \ldots, A_M, C\} & s_1 &= \{\neg A_1, A_2, \ldots, A_M, \neg C\} \\ s_2 &= \{A_1, \neg A_2, \ldots, A_M, \neg C\} & &\ldots \\ s_M &= \{A_1, A_2, \ldots, \neg A_M, \neg C\} & s_{M+1} &= \{\neg A_1, \neg A_2, A_3, \ldots, A_M, \neg C\} \\ s_{M+2} &= \ldots & s_{2^M-1} &= \{\neg A_1, \neg A_2, \ldots, \neg A_M, \neg C\}. \end{aligned}$$

Define $\mathcal{S} := \{s_i | 0 \leq i \leq 2^M - 1\}$. Now, distance-based approaches identify the elements of $\mathcal{S}$ (i.e. consistent judgment sets) with a set of points in $\mathbb{R}^{M+1}$, in order to calculate the *distance* between each judgment set and the average of the group members' judgment. This subset of $\mathbb{R}^{M+1}$ is, for each $t \geq 0$, defined by

$$O_t := \{x \in \{0,1\}^M \times \{0,t\} | x_{M+1} = t \prod_{i=1}^{M} x_i\}.$$

O_t defines the set of admissible judgment vectors (see table 2): a positive judgment on a premise ("A_1") is naturally identified with a 1, a negative judgment ("$\neg A_2$") with a 0, and the judgment of the conclusion (the $M+1$-th entry of the vector) is determined by the judgments on the premises. [1]

By this rationale, we obtain a canonical isomorphism $M_t : \mathcal{S} \to O_t$, mapping judgment sets to their geometric representations, and vice versa:

$$(M_t(s_i))_j = \begin{cases} 1 & \text{if } (j \leq M) \wedge (A_j \in s_i) \\ t & \text{if } (j = M+1) \wedge (i = 0) \\ 0 & \text{otherwise.} \end{cases} \qquad (1)$$

1. Thus, x_{M+1}, the conclusion entry, is assigned either 0 or t – the point of introducing the parameter t consists in manipulating the relative weight of the judgment on the conclusion, compared to the judgments on the premises. See below.

For a set of N group members, we can now define a family of *distance-based aggregation procedures* $(D_t^{M,N})_{t\geq 0} : \mathcal{S}^N \to \{C, \neg C\}$ that map N consistent individual judgment sets to a group judgment on the conclusion C.[2] The idea is to calculate the *distance* between the geometric representations of the "average judgment" and each consistent judgment set, and to opt for the consistent judgment set that comes closest to the group average. More precisely, $D_t^{M,N}$ is defined by

$$D_t^{M,N}(s^{(1)}, \dots, s^{(N)}) := \begin{cases} C & \text{if } d_t^0 < \min_{1\leq i\leq 2^M-1} d_t^j \\ \neg C & \text{otherwise,} \end{cases} \tag{2}$$

where

$$d_t^i := \|M_t(s_i) - \frac{1}{N}\sum_{j=1}^{N} M_t(s^{(j)})\|_1, \tag{3}$$

takes the distance between $M_t(s_i)$ and $\frac{1}{N}\sum_{j=1}^{N} M_t(s^{(j)})$, by means of the 1-norm or Hamming distance (i.e., the componentwise sum of the absolute values of a vector).[3] In particular, $D_t^{M,N}$ opts for conclusion C if and only if the (unique) judgment set that is compatible with C – namely $s_0 = \{A_1, A_2, \dots, A_M, C\}$ – is closer to the average submission of the group members than any other consistent judgment set. As equations (2) and (3) indicate, the term "closer" is defined by means of the distance in an appropriately transformed space, where parameter t governs the transformation map. Informally spoken, *t expresses the extra weight assigned to the agents' judgment on the conclusion*, apart from the information on C that is contained in the agents' judgments on the premises.

Now, PBP and CBP are part of the $D_t^{M,N}$-continuum, namely as the two extremes of the spectrum. Also, $D_1^{M,N}$ agrees with MFO (all proofs are in the appendix):

Proposition 1

$$D_t^{M,N} = \begin{cases} \mathit{PBP} & \textit{if } t = 0 \\ \mathit{MFO} & \textit{if } t = 1 \\ \mathit{CBP} & \textit{if } t = \infty. \end{cases}$$

In other words, the lower t, the more $D_t^{M,N}$ resembles the premise-based procedure because the judgments on the conclusion carry little weight, compared to the judgments on the premises. Vice versa for $t \to \infty$. In the remainder of the paper, we discuss the properties of the $D_t^{M,N}$-family, with special emphasis on three particular procedures: $D_0^{M,N}$ (the PBP), $D_1^{M,N}$ (the MFO) and $D_\infty^{M,N}$ (the CBP).

2. We differ from Miller and Osherson's [13] distance-based aggregation methods in several respects, most notably our specific parametrization, and our focus on the reliability of a method.

3. Note that in the group average, the $M+1$-th component is also parametrized by t, in order to express the relative weight which the judgments on the conclusion obtain.

3 Distance-Based Procedures: A Comparison

3.1 The Model

To compare the reliability of the distance-based procedures, we adopt a probabilistic framework. In particular, each agent is assigned an individual competence $p \in (0,1)$ to make a correct judgment about a single premise, independent of whether A_i or its negation $\neg A_i$ is true. In other words, the members' judgments on each premise are treated as independent random variables that mirror the truth with a certain probability p.[4] We could, for instance, imagine that the council members are laymen in the subject matter, have large files on both premises (supply/demand analysis and sustainability) and their task consists in evaluating these data without being misled.

Under suitable assumptions, the Condorcet Jury Theorem links the competence of the agents to the reliability of majority voting: if individual agents are better than randomizers at judging the truth or falsity of a premise (in other words, if $p > 0.5$), and if they form their opinions independently, then majority voting eventually yields the right collective judgment on A_i with increasing size of the group. This general fact motivates the use of majority-based decision-making in the judgment aggregation problem, e.g. in PBP and CBP.

Now we make the assumptions of our model explicit. They are also required to avoid computational complexity (cf. [1]):

(i) The marginal probabilities of the premises are equal ($\mathbb{P}(A_i) = \mathbb{P}(A_j)$).

(ii) The premises are (logically and probabilistically) independent.

(ii) All agents have the same (independent) competence to assess the truth of each single premise ($= p$). Their judgments on the premises are independent.

(iv) Each individual judgment set is logically consistent.

Assumptions (i) and (ii) entail that we can parametrize the set of prior distributions by a single parameter $q := \mathbb{P}(A_i) = \mathbb{P}(A_j)$. Then, we can define the reliability $R_{p,q}(D_t^{M,N})$ of a distance-based procedure for given parameter values N, p, q and t as the probability that the right conclusion is selected:[5]

$$R_{p,q}(D_t^{M,N}) \quad := \quad q^M \, \mathbb{P}(D_t^{M,N} = C|C) + (1-q^M)\, \mathbb{P}(D_t^{M,N} = \neg C|\neg C). \quad (4)$$

3.2 Reliability of the Distance-Based Procedures

With equation (4) at hand, we can compare the reliability of various distance-based procedures.

4. It would, of course, also be possible to assign two different competences p' and p'' to the agents, one for correctly discerning A_i and one for correctly discerning $\neg A_i$. But then, the agents' overall competence would be coupled to the prior probabilities of the various situations: $p = p' \, \mathbb{P}(A_i) + p'' \, (1 - \mathbb{P}(A_i))$ ($\mathbb{P}(A_i)$ denoting the marginal probability of A_i). – We ascribe an individual competence only for voting on *premises*, not for voting on any proposition (such as $A_i \wedge A_j$). Indeed, in many contexts, such as the harbor example, it is apparently reasonable to assign individual voting competence only to judgments on "elementary", matter-of-fact propositions, and not to complex propositions.

5. Here and in the sequel, we do not give the details of our calculations, apart from the proofs of the propositions which are contained in the appendix. These details can be readily provided, but they cost lots of space, and we doubt that they lead to any important insights.

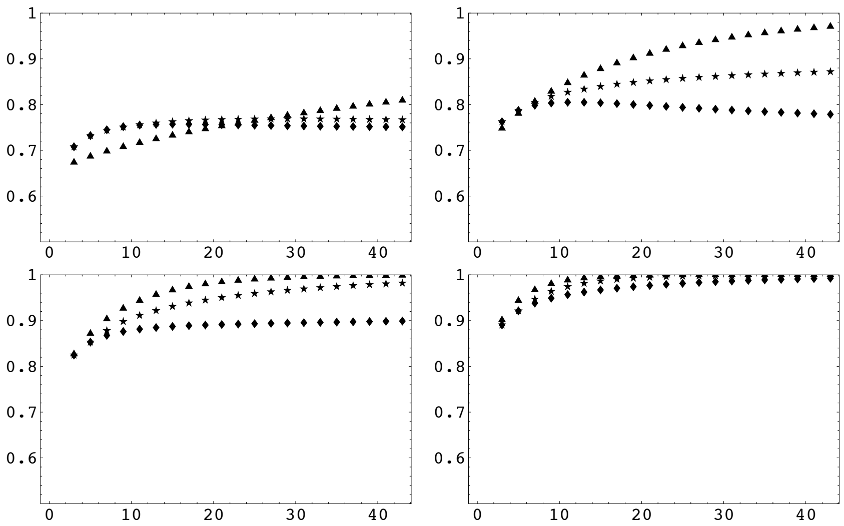

Figure 1: Reliability of PBP ($t = 0$, triangles), MFO ($t = 1$, stars) and CBP ($t = \infty$, diamonds) as a function of N, for $M = 2$, various values of p, and a fixed value of $q = 0.5$. Upper left figure: $p = 0.56$. Upper right figure: $p = 0.64$. Lower left figure: $p = 0.72$. Lower right figure $p = 0.8$.

Figure 1 depicts the reliability of distance-based procedures for two premises as a function of group size N, for two premises, $q = 0.5$ and individual competence $p \in \{0.56, 0.64, 0.72, 0.8\}$.[6] It turns out that for small groups and relatively low individual competence, the PBP too often leans towards an erroneous judgment in favor of C. In such small groups, majorities for a premise can emerge by random sampling effects alone and are therefore not informative. Therefore PBP is inferior to procedures with a higher t in such circumstances. However, if either p or q goes up (we don't discuss the latter case, but the argument is obvious), the three procedures either do not differ much (e.g. $p = 0.8$), or procedures with a low t outperform their competitors (e.g. $p = 0.64$). Thus, from an epistemic perspective, voting on *reasons* (premises) is clearly superior to directly voting on the conclusion as long as individual competence p, marginal probability q and group size N are not too low.

This insight can be accounted for, and generalized, by an asymptotic analysis $N \to \infty$: If individual competence exceeds a certain threshold, the group judgment on the conclusion will almost certainly be right as group size increases. First we show this for the case that the conclusion is actually true.

Proposition 2 *Assume $s_0 = \{A_1, A_2, \ldots, A_M, C\}$ is the true judgment set. Define*

$$p_t := \frac{\sqrt{2t^2 + 2t + 1} - 1}{2t}.$$

Then:

$$\lim_{N\to\infty} \mathbb{P}(D_t^{M,N} = C) = \begin{cases} 1 & \text{if } p > p_t \\ 0 & \text{if } p \le p_t. \end{cases}$$

6. For reasons of convenience, the calculations were made for two premises only, but our asymptotical results (Proposition 2, Corollary 1) ensure that the *structure* of the graphs is preserved for $M > 2$, too.

In particular, we obtain the critical values $p_0 = 0.5$ for PBP, $p_1 = (\sqrt{5}-1)/2 \approx 0.618$ for MFO and $p_\infty = 1/\sqrt{2} \approx 0.707$ for CBP.

The result easily transfers to the general case:

Corollary 1 *Let $p > 0.5$. Then $\mathbb{P}$-a.s.:*

$$\lim_{N\to\infty} R_{p,q}(D_t^{M,N}) \quad = \quad \begin{cases} 1 & p_t > p \\ 1 - q^M & \textit{otherwise.} \end{cases} \tag{5}$$

Put another way, $R_{p,q}(D_t^{M,N}) \overset{N\to\infty}{\to} 1$ if p is larger than the critical value p_t, and $R_{p,q}(D_t^{M,N}) \overset{N\to\infty}{\to} 1 - q^M$ otherwise.[7] Thus, Proposition 1 and Corollary 1 explain why in Figure 1, the reliability of each procedure approaches $1 - q^2 = 0.75$ with increasing N when $p_t \le p$ (such as in the upper figures). This result suggests that for medium to large size groups and unknown competence, it is preferable to confront the agents with a couple of minor questions, rather than with one big aggregate question which requires the combination of several independent judgments. A complex issue, such as a motion in parliament, or a plebiscite question, should, if possible, be split into several smaller issues. For instance, if we ask the members of a city council or the participants of a poll whether they consider the harbor project eligible for public funding, the epistemic significance of their judgments is increased if we derive the aggregate judgment from their judgments on the premises *and* the conclusion, rather than directly asking them what they think about the conclusion. Note that this effect does not stem from a supposed tendency to commit logical fallacies when combining several judgments. Rather, procedures with high t give extraordinary weight to the agents' opinion on the conclusion. Then, agents mostly lean towards $\neg C$ even if each premise is widely accepted. This systematic bias deteriorates the results unless the agents' individual competence is very high.

3.3 Generalizations

We chose quite special and restrictive model assumptions in order to control computational complexity. Therefore we also explore whether our results are robust when relaxing two assumptions: (i) the loss of a "false positive" decision (for C) exceeds the loss of a "false negative" decision (for $\neg C$), (ii) the independence of the premises is abandoned.

Firstly, quite often a judgment leads to an irreversible decision. For instance, if the newly built harbor happens to affect the marine reserve, or if the demands for new harbor sites are too low, we will continuously suffer from the consequences of a wrong judgment. (Whereas if we do not approve of the project, we are free to correct our judgment later.) In such a case, use of a ("premise-friendly") procedure with a low t is apparently dangerous. To deal with that case, we have set up a utility matrix where each correct decision is rewarded with 1 unit, a false rejection of C costs 0 units, and a false acceptance of C costs u units, with $u < 0$.

Figure 2 plots the expected utility of a procedure as a function of t for $u = -0.5$ and $u = -1$, with fairly typical values of p, q and N. We observe that

7. It is not surprising, by the way, that q enters the formula, because it expresses which form of systematic bias (towards C or $\neg C$) is more important for the reliability of $D_t^{M,N}$.

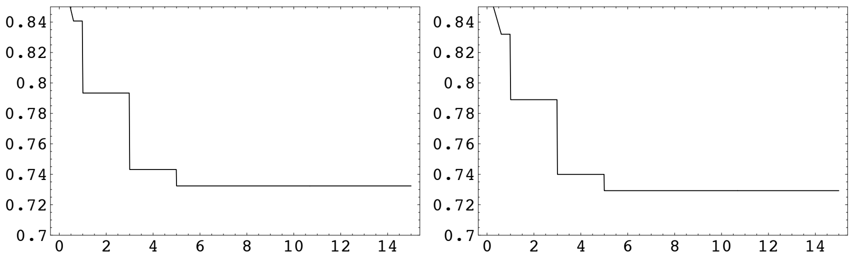

Figure 2: Expected utility of the distance-based procedures as a function of t, for $M = 2$, $N = 11$, $p = q = 0.7$. Left figure: $u = -0.5$. Right figure: $u = -1$.

the graph is monotonously decreasing as a function of t. In other words, even if the costs of false acceptance are significantly higher than the costs of false rejection, there is no need to employ a high t in aggregating the judgments.

Secondly, although in the harbor example, independence of the premises is quite plausible, it seems to be too strong an assumption in other contexts, e.g. when both premises express purely economic criteria. Consequently, the agents' judgments will quite often be far from independent, e.g. they might reason from the conclusion to the premises instead the other way round. This prompts the question how our distance-based procedures fare in such cases. For reasons of simplicity and computational tractability, we restrict ourselves to two premises in extending the model.

The occurrence of such a correlation seems to question the superiority of distance-based procedures with low t, e.g. the premise-based procedure, since the additional information obtained by voting on two premises, instead of voting on the conclusion, is low. To check this prediction, we set up a Bayes net with a parent node C, $P(C) = c$, and the premises A_1 and A_2 as offsprings. This creates a dependence between A_1 and A_2, expressed by $x = P(A_1|C) = 1 - P(A_1|\neg C)$, and analogously for A_2. Leaving the rest of our model assumptions intact, we have plotted the reliability of the distance-based procedures as a function of the strength of the correlation involved ($x = 0.5$ corresponds to independence, $x = 0$ or $x = 1$ to perfect correlation), for typical values of p and $N = 7$.

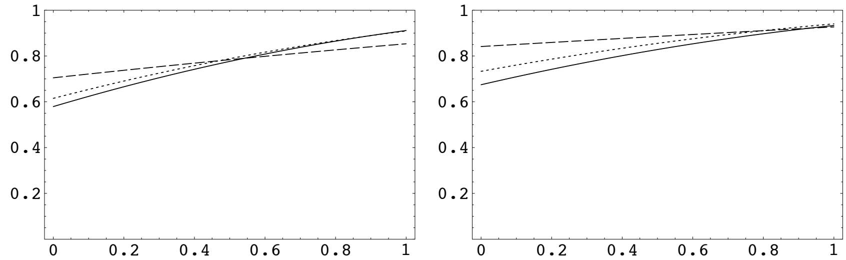

Figure 3: Reliability of PBP ($t = 0$, dashed line), MFO ($t = 0$, dotted line) and CBP ($t = \infty$, full line) as a function of x, for $M = 2$, $N = 7$, $c = 0.3$, and $p = 3/5$ (left figure) compared to $p = 2/3$ (right figure).

To better understand Figure 3, note that the marginal probabilities of A_1

and A_2 can be written as

$$P(A_1) = P(A_2) = cx + (1-c)(1-x),$$

which entails the probability of A_1 and A_2 decreases linearly as a function of x, with slope -0.4. This means that the apparently strong performance of CBP and MFO in the left figure, for $x > 0.5$, owes more to the decreasing probability of A_1 (which favors high ts), than to the increasing correlation. Once this effect is eliminated, the distortions induced by premise correlation are very modest. To sum up, while low t's apparently dominate higher t's for high individual competence ($p = 2/3$), for lower competences ($p = 3/5$) *slightly* higher t's, e.g. $t \approx 1$ are advised. This agrees with our findings from the previous subsection which remain thus intact for both kinds of generalizations.

4 A Group-External Perspective

The distance-based procedures aggregated a N-tuple of individual judgment sets to a final judgment on the conclusion. Potentially available additional information was neglected. Our approach in this section addresses this issue: An *external* decision-maker treats the agents' judgments on premises and conclusion as *evidence* for a final judgment on the conclusion. For instance, think of a prime minister, a city mayor, etc. who takes into account the recommendations of an advisory board, while having some opinion on the subject matter and the competence of the board members. We discuss the reliability of this kind of judgment aggregation as a function of the accuracy of the estimates of individual competence, and compare it to the distance-based procedures.

First, let us see how this approach can serve as a benchmark for distance-based procedures. The judgments of the agents $(s^{(1)}, \ldots, s^{(N)})$ are treated as incoming evidence and used to updates prior probabilities to a posterior distribution over $\{C, \neg C\}$. We obtain the formula

$$\mathbb{P}(C|(s^{(1)}, \ldots, s^{(N)})) \quad = \quad \left(1 + \frac{1 - \mathbb{P}(C)}{\mathbb{P}(C)} \frac{\mathbb{P}(s^{(1)}, \ldots, s^{(N)}|\neg C)}{\mathbb{P}(s^{(1)}, \ldots, s^{(N)})|C)}\right)^{-1}. \qquad (6)$$

Thus, the optimal decision is exclusively based on the posterior probability of C and the utility matrix which describes the decision problem. If both kinds of error (wrong judgment in favor of/against C) are equally severe, a rational Bayesian decision-maker will opt for C if and only if $\mathbb{P}(C|(s^{(1)}, \ldots, s^{(N)})) > 1/2$.

If we add the assumptions (i)-(iv) (see page 6) that served to determine the reliability of the distance-based procedures, the model is fully specified, and the right hand side of (6) can be readily calculated. In particular, we can compare the reliability of distance-based aggregation to the theoretically optimal Bayesian aggregation. Are they close to the theoretical optimum or do they suffer high losses?

Figure 4 gives an answer for typical values of q and N. As predicted, the Bayesian procedure constitutes an upper bound for all other procedures. Also, the more the agents are inclined towards either truth or falsehood, the more informative are their judgments for an external decision-maker. Hence the symmetry around $p = 0.5$. The most notable thing is, however, that the performance of PBP – representative of procedures that use a low t – comes very close to the

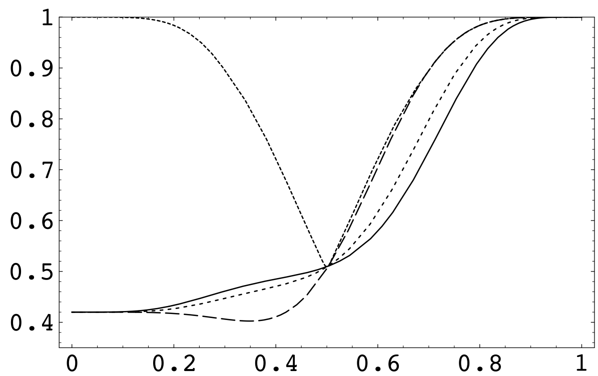

Figure 4: The probability that the Bayesian procedure (thin-dotted line) identifies the right conclusion as a function of the competence of the agents p for $q = .7$, $M = 2$ and $N = 11$, compared to distance-based procedures for $t = 0$ (dashed line), $t = 1$ (thick-dotted line) and $t = \infty$ (full line).

theoretically optimal Bayesian procedure. In other words: Although the PBP is a very crude procedure, we can, for typical values of q and N, barely improve upon it, even as external decision-makers with full information on p and q. This gives another justification for using easily implementable distance-based procedures in practice, and justifies rules of thumbs, such as "discuss and vote on the *criteria* for a decision, instead on the decision itself". For instance, proposals that are opposed by a majority of agents should not be dismissed when the criteria for their implementation are backed by strong majorities.

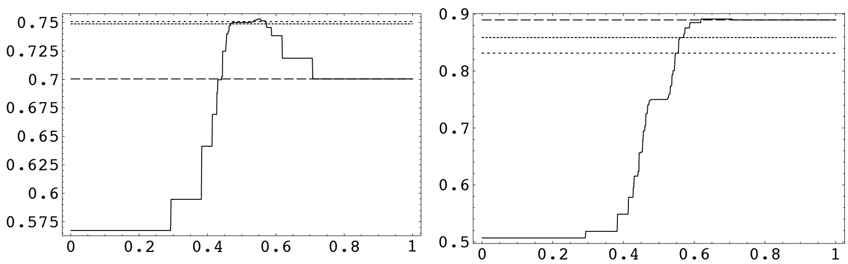

Figure 5: Reliability of the Bayesian procedure (full line) as a function of estimated competence $\hat{p}$, with $M = 2$, $N = 11$, $q = 0.5$, $p = 0.55$ (left figure) and $p = 2/3$ (right figure). The horizontal lines give the distance-based procedures for $t = 0$ (dashed line), $t = 1$ (thick-dotted line) and $t = \infty$ (thin-dotted line).

Of course, the pure Bayesian procedure is hard to implement in practice because it demands that p and q be transparent to the decision-maker. Therefore we investigate it under more realistic circumstances when the external decision-maker does *not* know the agents' individual competence and has to estimate it instead. Figure 5 plots the reliability of this more realistic form of Bayesian judgment aggregation and compares it to the performance of the distance-based procedures.

We see that the graphs of the distance-based procedures are constant – which is intuitively clear since they do not depend on the estimate $\hat{p}$. As expected, the Bayesian procedure peaks around the true value of p, outperforming the distance-based procedures in this region. For overestimated $\hat{p}$, it is at least as good as the PBP ($t = 0$), securing reasonable reliability. This is no surprise since if the agents are deemed extremely competent, the decision-maker will follow their majority judgments, mimicking PBP. By contrast, underestimating p leads to very unreliable conclusions.

This yields the following, more general moral: when external decision-makers have well-founded prior beliefs on the likelihood of the premises, and have to assess recommendations of an advisory board, to overestimate the board's competence is a safer strategy than to underestimate it. This might explain why reports of such boards (e.g. the IPCC report on global warming, or the Stern report on the economics of climate change) are often treated like revealed truths: even if we have doubts that the board members are competent enough to give binding recommendations on such complex a matter, *assuming* that they are competent is a more robust, and rational strategy than to systemically underestimate their competence, and to neglect their advice. Individual bias can just be too devastating.

5 Conclusions

This paper has investigated some popular judgment aggregation procedures (such as PBP, CBP and MFO) and embedded them into a more general family of distance-based procedures. This family has been parametrized by means of a real number $t \geq 0$, where the extremes are given by the premise- and the conclusion-based procedure.

We have proven asymptotic results for increasing group size and shown that unless the group is quite small and the voters barely competent, distance-based procedures with low t, such as the premise-based procedure, fare best. The Bayesian analysis has moreover demonstrated that these procedures fare almost as well as the theoretical optimum. Also, the results are robust with respect to two plausible generalizations, namely changing utility matrices and correlation among the premises. Although in real aggregation problems, the precise choice of an aggregation procedure has to be calibrated to the group size and problem specifics, as explained in section 3, our results substantiate a general recommendation: namely the use of procedures that give a low weight to the initial judgments on the conclusion, in particular the premise-based procedure PBP. These distance-based approaches are easily implementable and reliably select the right outcome without requiring judgments on individual competence.

Finally, we have adopted the perspective of an external decision-maker who merges the recommendation of an advisory board with his own beliefs on the subject matter and the board members' competence. The approach is very flexible since the decision procedure can be adapted to a lot of different problems (by changing the utility matrix). However, the board members' competence has typically to be estimated. It has been shown that for incorrect estimates of p, losses are manageable in the case of overestimation, whereas underestimation often proves to be fatal. This result encourages policy-makers to rely on the advice of an unbiased advisory board or committee, even if the majorities are not convincing and the board members' competence is hard to elicit.

Appendix

Lemma 1: Let d_t^j be defined according to equation (3). Then:

$$d_t^0 < min_{1 \leq j \leq 2^M - 1}\, d_t^j \qquad \Leftrightarrow \qquad d_t^0 < min_{1 \leq j \leq M}\, d_t^j.$$

Proof of Lemma 1: '$\Rightarrow$' is trivial. For the converse direction, assume $d_t^0 < min_{1 \leq j \leq M} d_t^j$ and the existence of a $K > M$ such that $d_t^0 \geq d_t^K$. We conduct an indirect proof and lead this assumption ad absurdum. Assume without loss of generality that $s_K = \{\neg A_1, \neg A_2, \ldots, \neg A_L, A_{L+1}, \ldots, A_M, C\}$, with $L \geq 2$ (the A_j are interchangeable). Let a_j, for any judgment set in $\mathcal{S}^N$, denote the actual number of votes for premise A_j, and accordingly for c with respect to conclusion C. Now, we rewrite the (1-norm) distances d_t^j between any $M_t(s_j)$ and the group average $\frac{1}{N}\sum_{i=1}^N M_t(s^{(i)})$:

$$d_t^0 = \frac{1}{N}\left[\sum_{j=1}^{N}(N - a_j) + t(N - c)\right] \qquad d_t^1 = \frac{1}{N}\left[\sum_{j \neq 1}(N - a_j) + a_1 + tc\right]$$

$$d_t^2 = \frac{1}{N}\left[\sum_{j \neq 2}(N - a_j) + a_2 + tc\right] \qquad d_t^K = \frac{1}{N}\left[\sum_{j \leq L} a_j + \sum_{j > L}(N - a_j) + tc\right].$$

This yields for all $1 \leq j \leq M$

$$N(d_t^j - d_t^0) = (2a_j - N) + t(2c - N). \tag{7}$$

Hence $d_t^j > d_t^0$ presupposes $(2a_j - N) > 0$ because $c \leq a_i$. – Finally we obtain

$$\begin{aligned} N(d_t^K - d_t^1) &= \sum_{j \leq L} a_j + \sum_{j=2}^{L}(N - a_j) - a_1 \\ &= \sum_{j=2}^{L} a_j - (N - a_j) \\ &= \sum_{j=2}^{L} 2a_j - N \\ &> 0, \end{aligned} \tag{8}$$

making use of (7) and the remark thereafter. – (8) contradicts the assumption that $d_t^K \leq d_t^0 < d_t^1$. □

Proof of Proposition 1: $D_1^{M,N}$ = MFO is obvious (compare table 2). We show that $D_0^{M,N}$ = PBP and $D_\infty^{M,N}$ = CBP. Recall that $D_t^{N,M} = C$ if and only if $d_t^0 < \min_{i \neq 0} d_t^i$ (cf. equation (2)). Lemma 1 has demonstrated that this condition is equivalent to $d_t^0 < \min_{1 \leq i \leq M} d_t^i$. In other words, $D_t^{M,N} = C$ if and only if

$$(2a_j - N) + t(2c - N) > 0 \qquad \forall 1 \leq j \leq M \tag{9}$$

(cf. (7)).

For $t = 0$, (9) can be rewritten as $a_j > N/2$ for each $1 \leq j \leq M$, and we obtain the PBP. For $t \to \infty$, (9) amounts to $c > N/2$, and we obtain the CBP. □

Proof of Proposition 2: Our proof is straightforward, using limit results from probability theory. By assumption we know that s_0 is the true situation. Then, the number of judgments in favor of a premise A_j, a_j, is a Binomially distributed random variable with parameter p and sample size N ($a \sim B_{N,p}$). Similarly, $c \sim B_{N,p^2}$. All variables can be written as sums of independent and identically distributed random variables so that the Strong Laws of Large Numbers applies. It follows that $\mathbb{P}$-a.s. $a_j/N, \stackrel{N\to\infty}{\longrightarrow} p$, $c/N \stackrel{N\to\infty}{\longrightarrow} p^2$. In particular, with the exception of a set of measure zero,

$$\forall\varepsilon \geq 0 \; \exists N_0 \; \forall N \geq N_0 : \; \frac{a_j}{N} \in (p-\varepsilon, p+\varepsilon), \; \frac{c}{N} \in (p^2-\varepsilon, p^2+\varepsilon). \tag{10}$$

Recall that $D_t^{M,N} = C$ if and only if

$$(2a_j - N) + t(2c - N) > 0 \qquad \forall 1 \leq j \leq M$$

(cf. (9)), which is equivalent to

$$2\frac{a_j}{N} - 1 + t\left(2\frac{c}{N} - 1\right) > 0 \qquad \forall 1 \leq j \leq M.$$

Now we show that this happens $\mathbb{P}$-almost surely if $p > p_t := (1/2t)(\sqrt{2t^2+2t+1} - 1)$ and $N \to \infty$. Let $(3+3t)\varepsilon := \left|2tp^2 + 2p - (1+t)\right|$. A simple computation yields that $2tp^2 + 2p - (1+t) > 0$ if and only if $p > p_t$ for $t \neq 0$ and $p > 0.5$ for $t = 0$. Hence, for $p > p_t$ and $N \geq N_0$, and any $1 \leq j \leq M$,

$$\begin{aligned} & 2\frac{a_j}{N} - 1 + t\left(2\frac{c}{N} - 1\right) \\ > \; & 2p - 2\varepsilon - 1 + t(2p^2 - 2\varepsilon - 1) \\ = \; & (3+3t)\varepsilon - 2\varepsilon - 2t\varepsilon \\ > \; & 0, \end{aligned}$$

using the limit result (10). Thus, if $p > p_t$, $\mathbb{P}$-almost surely $D_t^{M,N} = C$, as desired. Assume on the other hand that $p \leq p_t$. Then, in virtue of (10), for $N \geq N_0$,

$$\begin{aligned} & 2\frac{a_j}{N} - 1 + t\left(2\frac{c}{N} - 1\right) \\ < \; & 2p + 2\varepsilon - 1 + t(2p^2 + 2\varepsilon - 1) \\ = \; & (-3-3t)\varepsilon + 2\varepsilon + 2t\varepsilon \\ < \; & 0. \end{aligned}$$

This entails that, incorrectly, $D_t^{M,N} = \neg C$ $\mathbb{P}$-a.s. for increasing N. Thus, if $p > p_t$, the right conclusion will be eventually selected, and if $p \leq p_t$, the wrong conclusion will be eventually selected as $N \to \infty$.

It remains to calculate the critical values p_t for $t \in \{0, 1, \infty\}$. For $t \to \infty$ (CBP), a simple asymptotic analysis yields the condition $p > 1/\sqrt{2}$, whereas $t = 0$ (PBP) amounts to $p > 0.5$. Finally, in the case $t = 1$ (MFO), we get the threshold $p > (\sqrt{5} - 1)/2$. $\square$

Proof of Corollary 1: Proposition 2 has investigated the reliability of $D_t^{M,N}$ for the case that C is true. Now we presuppose that C is false. In that case, there will, for each $p > 0.5$ and increasing N, be an a_j such that $\mathbb{P}$-a.s. $a_j < N/2$. This implies that $D_t^{M,N} = \neg C$, independent of t (cf. (9)). Hence all distance-based procedures $D_t^{M,N}$ eventually perform reliably if C is false. Combining this result with proposition 2, we obtain

$$\begin{aligned} R_{p,q}(D_t^{M,N}) &= \mathbb{P}(C)\mathbb{P}(D_t^{M,N} = C|C) + \mathbb{P}(\neg C)\mathbb{P}(D_t^{M,N} = \neg C|\neg C) \\ &= q^M \, 1_{\{p_t > p\}} + (1 - q^M), \end{aligned}$$

proving equation (5). $\square$

References

[1] L. Bovens and W. Rabinowicz. Democratic answers to complex questions. An epistemic perspective, *Synthese* 150: 131–153, 2006.

[2] B. Chapman. Rational aggregation. *Politics, Philosophy and Economics*, 1(3): 337-354, 2002.

[3] F. Dietrich. Judgment aggregation: (im)possibility theorems, *Journal of Economic Theory* 126: 286–298. 2006.

[4] F. Dietrich and C. List. Arrow's theorem in judgment aggregation. *Social Choice and Welfare* 29: 19–33, 2007.

[5] S. Konieczny and E. Grégoire. Logic-based approaches to information fusion. *Information Fusion* 7: 4–18, 2006.

[6] S. Konieczny and R. Pino-Pérez. Merging with integrity constraints. In *Fifth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'99)*, 233–244, 1999.

[7] L.A. Kornhauser and L.G. Sager. Unpacking the court. *Yale Law Journal*, 96: 82-117, 1986.

[8] L.A. Kornhauser and L.G. Sager. The one and the many: Adjudication in collegial courts. *California Law Review*, 81: 1-51, 1993.

[9] C. List. The probability of inconsistencies in complex collective decisions. *Social Choice and Welfare* 24: 3–32, 2005.

[10] C. List and R. Goodin. Epistemic Democracy: Generalizing the Condorcet Jury Theorem. *Journal of Political Philosophy* 9: 277–306, 2001.

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

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

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

[14] P. Pettit. Deliberative democracy and the discursive dilemma. *Philosophical Issues*, 11: 268-299, 2001.

[15] G. Pigozzi. Belief merging and the discursive dilemma: an argument-based account to paradoxes of judgment aggregation. *Synthese* 152: 285–298, 2006.