Manipulation in group argument evaluation (Extended Abstract)

Martin Caminada
Individual and Collective Reasoning, University of Luxembourg
martin.caminada@uni.lu

Gabriella Pigozzi
LAMSADE
Université Paris-Dauphine
France
gabriella.pigozzi@dauphine.fr

Mikołaj Podlaszewski
Individual and Collective Reasoning, University of Luxembourg
mikolaj.podlaszewski@gmail.com

ABSTRACT

Given an argumentation framework and a group of agents, the individuals may have divergent opinions on the status of the arguments. If the group needs to reach a common position on the argumentation framework, the question is how the individual evaluations can be mapped into a collective one. This problem has been recently investigated in [1]. In this paper, we study under which conditions these operators are Pareto optimal and whether they are manipulable.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*multiagent systems*

General Terms

Economics, Theory

Keywords

Collective decision making, Argumentation, Judgment aggregation, Social choice theory

1. INTRODUCTION

Individuals can hold different reasonable positions on the information they share. In this paper we are interested in group decisions where members share the same information. One of the principles of argumentation theory is that an argumentation framework can have several extensions/labellings. If the information the group shares is represented by an argumentation framework, and each agent's reasonable position is an extension/labelling of that argumentation framework, the question is how to aggregate the individual positions into a collective one.

Caminada and Pigozzi [1] have studied this issue in abstract argumentation and provided three aggregation operators. The key property of these operators is that the collective outcome is 'compatible' with each individual position. That is, an agent who has to defend the collective position in public will never have to argue directly against his own private position.

Cite as: Manipulation in group argument evaluation (Extended Abstract), Martin Caminada, Gabriella Pigozzi and Mikołaj Podlaszewski, *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, Tumer, Yolum, Sonenberg and Stone (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. 1127-1128.
Copyright © 2011, International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

In this paper we focus on the behaviour of two of the three aggregation operators of [1] and address the following research questions:

(i) Are the social outcomes of the aggregation operators in [1] Pareto optimal if preferences between different outcomes are also taken into account?

(ii) Do agents have an incentive to misrepresent their own opinion in order to obtain a more favourable outcome? And what are the effects from the perspective of social welfare?

Due to page constraints, we refer the reader to [1] for an outline of abstract argumentation theory and for the definitions of the sceptical and credulous aggregation operators.

2. PREFERENCES

In order to investigate Pareto optimality and strategy-proofness we need to assume that agents have preferences over the possible collective outcomes. We write $\mathcal{L} \geq_i \mathcal{L}'$ to denote that agent i *prefers* labelling $\mathcal{L}$ to $\mathcal{L}'$. We write $\mathcal{L} \sim_i \mathcal{L}'$, and say that i *is indifferent* between $\mathcal{L}$ and $\mathcal{L}'$, iff $\mathcal{L} \geq_i \mathcal{L}'$ and $\mathcal{L}' \geq_i \mathcal{L}$. Finally, we write $\mathcal{L} >_i \mathcal{L}'$ (agent i *strictly prefers* $\mathcal{L}$ to $\mathcal{L}'$) iff $\mathcal{L} \geq_i \mathcal{L}'$ and not $\mathcal{L} \sim_i \mathcal{L}'$.

We assume that the labelling submitted by each agent is his most preferred one and, hence, the one he would like to see adopted by the whole group. The order over the other possible labellings is generated according to the distance from the most preferred one. For this purpose, we define Hamming sets and Hamming distance among labellings.

DEFINITION 1. *Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be two labellings of argumentation framework. We define the* Hamming set *between these labellings as $\mathcal{L}_1 \ominus \mathcal{L}_2 = \{A \mid \mathcal{L}_1(A) \neq \mathcal{L}_2(A)\}$ and the* Hamming distance *as $\mathcal{L}_1 \,|\ominus|\, \mathcal{L}_2 = |\mathcal{L}_1 \ominus \mathcal{L}_2|$.*

We are now ready to define an agent's preference given by the Hamming set and the Hamming distance as follows.

DEFINITION 2. *Let (Ar, def) be an argumentation framework, $\mathcal{L}abellings$ the set of all its labellings and $\geq_i$ the preference of agent i. We say that agent i's preference is* Hamming set based *(written as $\geq_{i,\ominus}$) iff $\forall \mathcal{L}, \mathcal{L}' \in \mathcal{L}abellings, \mathcal{L} \geq_i \mathcal{L}' \Leftrightarrow \mathcal{L} \ominus \mathcal{L}_i \subseteq \mathcal{L}' \ominus \mathcal{L}_i$ where $\mathcal{L}_i$ is the agent's most preferred labelling. Similarly, we say that agent i's preference is* Hamming distance based *(written as $\geq_{i,|\ominus|}$) iff $\forall \mathcal{L}, \mathcal{L}' \in \mathcal{L}abellings, \mathcal{L} \geq_i \mathcal{L}' \Leftrightarrow \mathcal{L} \,|\ominus|\, \mathcal{L}_i \leq \mathcal{L}' \,|\ominus|\, \mathcal{L}_i$ where $\mathcal{L}_i$ is the agent's most preferred labelling.*

We now have the machinery to represent individual preferences over the collective outcomes. We can now turn to the first research question of the paper, i.e., whether the sceptical and credulous aggregation operators are Pareto optimal.

	Sceptical Operator	Credulous Operator
Hamming set	Yes (Theorem 1)	Yes (Theorem 3)
Hamming dist.	Yes (Theorem 2)	No (Observation 1)

Table 1: Pareto optimality of the aggregation operators depending on the type of preference.

3. PARETO OPTIMALITY

Pareto optimality is a fundamental social welfare principle that guarantees that it is not possible to improve a social outcome, i.e. it is not possible to make one individual better off without making at least one other person worse off.

DEFINITION 3. *Let $N = 1, \ldots, n$ be a group of agents with preferences $\geq_i, i \in N$. $\mathcal{L}$* Pareto dominates *$\mathcal{L}'$ iff $\forall i \in N$, $\mathcal{L} \geq_i \mathcal{L}'$ and $\exists j \in N, \mathcal{L} >_j \mathcal{L}'$.*

A labelling is Pareto optimal if it is not dominated by any other labelling.

DEFINITION 4. *Labelling $\mathcal{L}$ is* Pareto optimal *if there is no $\mathcal{L}' \neq \mathcal{L}$ such that $\forall i \in N$, $\mathcal{L}' \geq_i \mathcal{L}$ and $\exists j \in N, \mathcal{L}' >_j \mathcal{L}$.*

We say that an aggregation operator is Pareto optimal if all its outcomes are Pareto optimal.

THEOREM 1. *If individual preferences are Hamming set based, then the sceptical aggregation operator is Pareto optimal when choosing from the admissible labellings that are smaller or equal (w.r.t $\sqsubseteq$) to each of the participants' individual labellings.*

THEOREM 2. *If individual preferences are Hamming distance based, then the sceptical aggregation operator is Pareto optimal when choosing from the admissible labellings that are smaller or equal (w.r.t $\sqsubseteq$) to each individual labellings.*

THEOREM 3. *If individual preferences are Hamming set based, then the credulous aggregation operator is Pareto optimal when choosing from the admissible labellings that are compatible ($\approx$) to each of the participants' labellings.*

OBSERVATION 1. *The credulous aggregation operator is not Pareto optimal when the preferences are Hamming distance based. This can be shown with an example, not included due to space constraints.*

We summarise our results in Table 1.

4. STRATEGIC MANIPULATION

When an agent knows the positions of the other agents, he may have an incentive to submit an insincere position. If an aggregation rule is manipulable, an agent may obtain a social outcome that is closer to his actual preferences by submitting an insincere input. Hence, it is important to study whether the aggregation operators are strategy-proof (i.e. non-manipulable). Profile $P_{\mathcal{L}_k/\mathcal{L}'_k}$ is profile P where agent k's labelling $\mathcal{L}_k$ has been changed to $\mathcal{L}'_k$.

DEFINITION 5. *Let P be a profile and $\mathcal{L}_k \in P$ the most preferred labelling of an agent with preference $\geq_k$. Let O be any aggregation operator. A labelling $\mathcal{L}'_k$ such that $O(P_{\mathcal{L}_k/\mathcal{L}'_k}) >_i O(P)$ is called a* strategic lie.

DEFINITION 6. *An aggregation operator O is* strategy-proof *if strategic lies are not possible.*

	Sceptical	Credulous
Hamm. set	No (Obs. 3) but benev. (Th. 4)	No and not benev. (Obs. 2)
Hamm. dist.	No (Obs. 3) but benev. (Th. 4)	No and not benev. (Obs. 2)

Table 2: Strategy-proofness of operators depending on the type of preference.

OBSERVATION 2. *The credulous aggregation operator is not strategy-proof (the example is omitted for space reasons).*

OBSERVATION 3. *The sceptical aggregation operator is not strategy-proof (the example is omitted for space reasons).*

Surprisingly, the lie under the sceptical operator does not harm the other agent. On the contrary, it improves the social outcome for both the agents. We call these lies *benevolent*.

THEOREM 4. *Under the sceptical aggregation operator and Hamming distance or Hamming set based preferences, for any agent, his strategic lies are benevolent.*

We summarise our results in Table 2.

5. CONCLUSION AND RELATED WORK

The study of aggregation problems in abstract argumentation is recent. For example, [2] presents an approach to merge Dung's argumentation frameworks.

Given an argumentation framework, [4] address the question of how to aggregate individual labellings into a collective position. By drawing on a general impossibility theorem from judgment aggregation, they prove an impossibility result and provide some escape solutions. Relevant for the present paper is another work by [3], where they explore welfare properties of collective argument evaluation.

In this paper we have analyzed the sceptical and credulous aggregation operators from a social welfare perspective. We have studied under which conditions these operators are Pareto optimal and whether they are manipulable. In future, we plan to consider focal set oriented agents, that is, agents who care only about a subset of the argumentation framework. We also plan to investigate distances that assign higher values to `in-out` conflicts than to `in-undec` or `out-undec`.

6. REFERENCES

[1] M. Caminada and G. Pigozzi. On judgment aggregation in abstract argumentation. *Autonomous Agents and Multi-Agent Systems*, 22(1):64–102, 2011.

[2] S. Coste-Marquis, C. Devred, S. Konieczny, M.-C. Lagasquie-Schiex, and P. Marquis. On the merging of dung's argumentation systems. *Artificial Intelligence*, 171(10-15):730–753, 2007.

[3] I. Rahwan and K. Larson. Welfare properties of argumentation-based semantics. In *Proceedings of the 2nd International Workshop on Computational Social Choice (COMSOC)*, 2008.

[4] I. Rahwan and F. Tohmé. Collective argument evaluation as judgment aggregation. In *Proc. of 9th AAMAS*, 2010.