Manipulation in Group Argument Evaluation

Martin Caminada*
SnT
Université du Luxembourg
Luxembourg
martin.caminada@uni.lu

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

Mikołaj Podlaszewski
ILIAS / SnT
Université du Luxembourg
Luxembourg
mikolaj.podlaszewski@uni.lu

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 by Caminada and Pigozzi. In this paper, we investigate the behaviour of two of such operators from a social choice-theoretic point of view. In particular, we study under which conditions these operators are Pareto optimal and whether they are manipulable.

1 Introduction

Individuals may draw different conclusions from the same information. Members of a jury may disagree on the verdict even though each member possesses the same information on the case under discussion. This happens because individuals can hold different reasonable positions on the information they share. Hence, the question is how the group can reach a common stance starting from the positions of each member.

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 [Caminada and Pigozzi, 2011] 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.

The aim of this paper is to formalise and examine the intuition that, although every social outcome that is compatible with one's own labelling is acceptable, some outcomes are more acceptable than others. That is, a collective outcome is more acceptable than another if it is compatible and more similar to one's own position than the other. In order to capture how much the various possible positions differ from each other, we use the notion of distance among labellings. Distance-based approaches have already been used to tackle aggregation problems, like in social choice theory [Brams *et al.*, 2007], belief merging [Konieczny and Perez, 1998] and its application to judgment aggregation [Pigozzi, 2006]. Thus, we say that a collective outcome is more acceptable than another if it is compatible, but the distance to one's own labelling is smaller than the other.

The observations above give rise to two new research questions, to be addressed in the current paper:

(i) Are the social outcomes of the aggregation operators in [Caminada and Pigozzi, 2011] 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 if so, what are the effects from the perspective of social welfare?

We focus on the behaviour of two of the three aggregation operators of [Caminada and Pigozzi, 2011]. Pareto optimality is a key principle of welfare economics which intuitively stipulates that a social state cannot be further improved. Thus, the first contribution of the paper is to study whether the compatible social outcomes selected by our aggregation operators are Pareto optimal. In order to investigate Pareto optimality, we consider the submitted labelling as the individual's most preferred option. By using a notion of distance, we derive the individual preference ordering over the other permissible labellings. We show that the two aggregation operators are Pareto optimal, when a certain distance is used.

The second contribution is on the manipulability of the aggregation operators. Manipulability is usually considered to be an undesirable property of social choice decision rules. If an aggregation rule is manipulable, an individual may, upon learning the preferences of the other agents, misrepresent his input to ensure a social outcome that is better for him than it would have been had he voted sincerely. Our findings show that, while the two operators are manipulable, the sceptical aggregation operator guarantees that an agent who lies does not only ensure a preferable outcome for himself, but even promotes social welfare, what we call a *benevolent lie*.

Section 2 outlines the abstract argumentation theory. In

*Supported by the National Research Fund, Luxembourg (LAAMI project).

Section 3 we define preferences over the individual evaluations. Pareto optimality and manipulability issues are addressed in Section 4 and 5 respectively. Section 6 concludes.

2 Preliminaries

2.1 Argumentation preliminaries

Definition 1 (Argumentation framework). *Let U be the universe of all possible arguments. An* argumentation framework *is a pair (Ar, def) where Ar is a finite subset of U and $def \subseteq Ar \times Ar$.*

We say that an argument A *defeats* (or *attacks*) an argument B iff $(A, B) \in def$. For example, in Fig. 1, we have that A attacks B and that B attacks C.

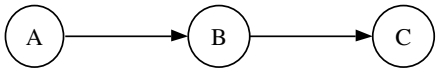

Figure 1: An argumentation framework.

Following [Caminada and Pigozzi, 2011], we use the argument labellings approach of [Caminada, 2006] rather than Dung's original extension approach [Dung, 1995]. The idea of a labelling is to associate with each argument exactly one label, which can either be `in`, `out` or `undec`. The label `in` indicates that the argument is explicitly accepted, `out` indicates that the argument is explicitly rejected, and `undec` indicates that one abstains from an explicit position on the argument.

Definition 2 (Labelling). *Let (Ar, def) be an argumentation framework. A* labelling *is a total function $\mathcal{L} : Ar \longrightarrow \{\texttt{in}, \texttt{out}, \texttt{undec}\}$.*

We write $\texttt{in}(\mathcal{L})$ for $\{A \mid \mathcal{L}(A) = \texttt{in}\}$, $\texttt{out}(\mathcal{L})$ for $\{A \mid \mathcal{L}(A) = \texttt{out}\}$ and $\texttt{undec}(\mathcal{L})$ for $\{A \mid \mathcal{L}(A) = \texttt{undec}\}$. Sometimes, we write a labelling $\mathcal{L}$ as a triple $(\mathcal{A}rgs_1, \mathcal{A}rgs_2, \mathcal{A}rgs_3)$ where $\mathcal{A}rgs_1 = \texttt{in}(\mathcal{L})$, $\mathcal{A}rgs_2 = \texttt{out}(\mathcal{L})$ and $\mathcal{A}rgs_3 = \texttt{undec}(\mathcal{L})$. When it only matters whether an agent has a clear position (`in` or `out`) on an argument, we write $\texttt{dec}(\mathcal{L})$ for $\texttt{in}(\mathcal{L}) \cup \texttt{out}(\mathcal{L})$.

Although labellings allow one to express any position on which arguments are accepted, rejected or left undecided, some of these positions are more reasonable than others. Basically, the task of an argumentation semantics is to provide a criterion for determining which positions can be considered to be reasonable. Several such semantics have been defined in the literature, but we will consider only the following ones.[1]

Definition 3 (Illegal arguments). *Let $\mathcal{L}$ be a labelling of argumentation framework (Ar, def) and let $A \in Ar$. We say:*

1. *A is* illegally `in` *iff A is labelled* `in` *but not all its defeaters are labelled* `out`
2. *A is* illegally `out` *iff A is labelled* `out` *but it does not have a defeater that is labelled* `in`
3. *A is* illegally `undec` *iff A is labelled* `undec` *but either all its defeaters are labelled* `out` *or it has a defeater that is labelled* `in`.

[1]For an explanation of why this is not a too restrictive assumption, the reader is referred to [Caminada and Pigozzi, 2011].

For example, argument A of Fig. 1 is `in` as it has no defeaters. Argument B must be `out` since it has a defeater (argument A) and its defeater is `in`. Finally, argument C is `in` since its only defeater (argument B) is `out`.

Definition 4 (Admissible labelling). *An* admissible labelling *is a labelling without arguments that are illegally* `in` *and without arguments that are illegally* `out`.

Definition 5 (Complete labelling). *A* complete labelling *is a labelling without arguments that are illegally* `in`*, without arguments that are illegally* `out` *and without arguments that are illegally* `undec`.

The basic difference between an extension [Dung, 1995] and a labelling is that an extension only represents the arguments that are accepted, whereas a labelling also represents the arguments that are rejected or left undecided. So labellings provide a slightly more expressive though equivalent way to express Dung's theory of argumentation. For results on how labellings relate to extensions, see [Caminada, 2006]

In essence, a labelling based semantics can be seen as a function that, given an argumentation framework, yields zero or more labellings, each of which can be seen as a reasonable position that one can take on an argumentation framework.

2.2 Aggregation problem

We are now ready to formally summarize the problem of aggregation of individual labellings into a collective position on a given argumentation framework. The definitions and results in this section are from [Caminada and Pigozzi, 2011].

Given a set of individuals $N = \{1, \ldots, n\}$, we need to define a general labellings aggregation operator O_{AF} that assigns a collective labelling $\mathcal{L}_{Coll}$ to each *profile* $P = \{\mathcal{L}_1, \ldots, \mathcal{L}_n\}$ of individual admissible labellings.

Definition 6 (Labelling aggregation operator O_{AF}). *Let $\mathcal{L}abellings$ be the set of all possible labellings of argumentation framework $AF = (Ar, def)$. A general labellings aggregation operator is a function $O_{AF} : 2^{\mathcal{L}abellings} - \{\emptyset\} \rightarrow \mathcal{L}abellings$ such that $O_{AF}(\{\mathcal{L}_1, \ldots, \mathcal{L}_n\}) = \mathcal{L}_{Coll}$.*

We are interested in aggregation operators that produce a collective outcome *compatible* with the individual opinions. The idea is to ensure that each member can publicly defend the common decision without having to directly go against his own position. Two notions of compatibility have been introduced in [Caminada and Pigozzi, 2011].

Definition 7 (Less or equally committed $\sqsubseteq$). *Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be two labellings of argumentation framework $AF = (Ar, def)$. We say that $\mathcal{L}_1$ is less or equally committed as $\mathcal{L}_2$ ($\mathcal{L}_1 \sqsubseteq \mathcal{L}_2$) iff $\texttt{in}(\mathcal{L}_1) \subseteq \texttt{in}(\mathcal{L}_2)$ and $\texttt{out}(\mathcal{L}_1) \subseteq \texttt{out}(\mathcal{L}_2)$.*

Definition 8 (Compatible labellings $\approx$). *Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be two labellings of argumentation framework (Ar, def). We say that $\mathcal{L}_1$ is* compatible *with $\mathcal{L}_2$ (denoted as $\mathcal{L}_1 \approx \mathcal{L}_2$) iff $\texttt{in}(\mathcal{L}_1) \cap \texttt{out}(\mathcal{L}_2) = \emptyset$ and $\texttt{out}(\mathcal{L}_1) \cap \texttt{in}(\mathcal{L}_2) = \emptyset$.*

The intuition is that, in order to be compatible, two labellings cannot have $\texttt{in} - \texttt{out}$ conflicts. It can be noted that $\sqsubseteq$ is a partial order on labellings, whereas $\approx$ is not transitive. It holds that if $\mathcal{L}_1 \sqsubseteq \mathcal{L}_2$, then $\mathcal{L}_1 \approx \mathcal{L}_2$. We are now ready to state the *sceptical* and *credulous* aggregation operators.

Definition 9 (Sceptical initial aggregation operator sio_{AF}). *Let $\mathcal{L}abellings$ be the set of all possible labellings of argumentation framework $AF = (Ar, def)$. The* sceptical initial aggregation operator *is a function $sio_{AF} : 2^{\mathcal{L}abellings} - \{\emptyset\} \rightarrow \mathcal{L}abellings$ such that $sio_{AF}(\{\mathcal{L}_1, \ldots, \mathcal{L}_n\}) =$*
$\{(A, \texttt{in}) \mid \forall i \in \{1, \ldots, n\} : \mathcal{L}_i(A) = \texttt{in}\} \cup$
$\{(A, \texttt{out}) \mid \forall i \in \{1, \ldots, n\} : \mathcal{L}_i(A) = \texttt{out}\} \cup$
$\{(A, \texttt{undec}) \mid \exists i \in \{1, \ldots, n\} : \mathcal{L}_i(A) \neq \texttt{in} \wedge \exists j \in \{1, \ldots, n\} : \mathcal{L}_j(A) \neq \texttt{out}\}$.

The idea is that the group initially labels an argument `in` (resp. `out`) if all individual participants agree that the argument is `in` (resp. `out`). Otherwise it is `undec`. This procedure does not preserve admissibility. This means that it may return a labelling with illegally `in` or illegally `out` arguments. This is why, after the initial aggregation, a second iterative phase follows, where all the illegally `in` or `out` arguments are relabelled to `undec`. Formally, this is defined as follows:

Definition 10 (Down-admissible labelling). *Let $\mathcal{L}$ be a labelling of argumentation framework $AF = (Ar, def)$. The* down-admissible *labelling of $\mathcal{L}$ is the biggest element of the set of all admissible labellings that is less or equally committed than $\mathcal{L}$.*

The down-admissible labelling is defined according to the partial order given by $\sqsubseteq$. Such element always exists and is unique [Caminada and Pigozzi, 2011]. We can now define the sceptical operator that ensures admissible outcomes.

Definition 11 (Sceptical aggregation operator so_{AF}). *Let $\mathcal{L}abellings$ be the set of all labellings of argumentation framework $AF = (Ar, def)$. The* sceptical aggregation operator *is a function $so_{AF} : 2^{\mathcal{L}abellings} - \{\emptyset\} \rightarrow \mathcal{L}abellings$ such that $so_{AF}(\{\mathcal{L}_1, \ldots, \mathcal{L}_n\})$ is the down-admissible labelling of $sio_{AF}(\{\mathcal{L}_1, \ldots, \mathcal{L}_n\})$.*

The aggregation operator above produces social outcomes that are less or equally committed to all the individual labellings. This result is also maximal:

Theorem 1. *Let $\mathcal{L}_1, \ldots, \mathcal{L}_n$ ($n \geq 1$) be labellings of argumentation framework $AF = (Ar, def)$. Let $\mathcal{L}_{so}$ be $so_{AF}(\{\mathcal{L}_1, \ldots, \mathcal{L}_n\})$. It holds that $\mathcal{L}_{so}$ is the biggest admissible labelling such that for every $i \in \{1, \ldots, n\}$: $\mathcal{L}_{so} \sqsubseteq \mathcal{L}_i$.*

The second aggregation operator is the *credulous* one.

Definition 12 (Credulous initial aggregation operator cio_{AF}). *Let $\mathcal{L}abellings$ be the set of all possible labellings of argumentation framework $AF = (Ar, def)$. The* credulous initial aggregation operator *is a function $cio_{AF} : 2^{\mathcal{L}abellings} - \{\emptyset\} \rightarrow \mathcal{L}abellings$ such that $cio_{AF}(\{\mathcal{L}_1, \ldots, \mathcal{L}_n\}) =$*
$\{(A, \texttt{in}) \mid \exists i \in \{1, \ldots, n\} : \mathcal{L}_i(A) = \texttt{in} \wedge \neg\exists j \in \{1, \ldots, n\} : \mathcal{L}_j(A) = \texttt{out}\} \cup$
$\{(A, \texttt{out}) \mid \exists i \in \{1, \ldots, n\} : \mathcal{L}_i(A) = \texttt{out} \wedge \neg\exists j \in \{1, \ldots, n\} : \mathcal{L}_j(A) = \texttt{in}\} \cup$
$\{(A, \texttt{undec}) \mid \forall i \in \{1, \ldots, n\} : \mathcal{L}_i(A) = \texttt{undec} \vee (\exists j \in \{1, \ldots, n\} : \mathcal{L}_j = \texttt{in} \wedge \exists k \in \{1, \ldots, n\} : \mathcal{L}_k = \texttt{out})\}$.

The idea is that the group initially labels an argument A `in` (resp. `out`) if there is someone who believes A is `in` (resp. `out`) and nobody thinks A is `out` (resp. `in`). A is labelled `undec` in all other cases. The admissibility problem reappears here and is solved again by an iterative second phase where all illegally `in` and `out` arguments are relabelled `undec`.

Definition 13 (Credulous aggregation operator co_{AF}). *Let $\mathcal{A}dm\mathcal{L}abellings$ be the set of all admissible labellings of argumentation framework $AF = (Ar, def)$. The credulous aggregation operator is a function $co_{AF} : 2^{\mathcal{A}dm\mathcal{L}abellings} - \{\emptyset\} \rightarrow \mathcal{A}dm\mathcal{L}abellings$ such that $co_{AF}(\{\mathcal{L}_1, \ldots, \mathcal{L}_n\})$ is the down-admissible labelling of $cio_{AF}(\{\mathcal{L}_1, \ldots, \mathcal{L}_n\})$.*

It holds that the credulous outcome labelling ($\mathcal{L}_{co} = co_{AF}(\{\mathcal{L}_1, \ldots, \mathcal{L}_n\})$) is compatible with all the individual labellings, i.e. $\mathcal{L}_{co} \approx \mathcal{L}_i$ (for each $i \in \{1, \ldots, n\}$).

3 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 now define Hamming sets and Hamming distance among labellings.

Definition 14 (Hamming set $\ominus$). *Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be two labellings of argumentation framework (Ar, def). We define the* Hamming set *between these labellings as $\mathcal{L}_1 \ominus \mathcal{L}_2 = \{A \mid \mathcal{L}_1(A) \neq \mathcal{L}_2(A)\}$.*

Definition 15 (Hamming distance $|\ominus|$). *Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be two labellings of argumentation framework (Ar, def). We define the* Hamming distance *between these labellings as $\mathcal{L}_1 \,|\ominus|\, \mathcal{L}_2 = |\mathcal{L}_1 \ominus \mathcal{L}_2|$.*

The Hamming set is the set of arguments on which two labellings differ, whereas the Hamming distance is the number of arguments on which two labellings differ. Since the labellings have only three values, we can use this lemma.

Lemma 1. *Let (Ar, def) be an argumentation framework and $\mathcal{L}_1$ and $\mathcal{L}_2$ two labellings:*

a) $\mathcal{L}_1 \ominus \mathcal{L}_2 = \texttt{in}(\mathcal{L}_1) \cap \texttt{out}(\mathcal{L}_2) \cup \texttt{in}(\mathcal{L}_1) \cap \texttt{undec}(\mathcal{L}_2) \cup \texttt{out}(\mathcal{L}_1) \cap \texttt{in}(\mathcal{L}_2) \cup \texttt{out}(\mathcal{L}_1) \cap \texttt{undec}(\mathcal{L}_2) \cup \texttt{undec}(\mathcal{L}_1) \cap \texttt{in}(\mathcal{L}_2) \cup \texttt{undec}(\mathcal{L}_1) \cap \texttt{out}(\mathcal{L}_2)$

b) if $\mathcal{L}_1 \sqsubseteq \mathcal{L}_2$ then $\mathcal{L}_1 \ominus \mathcal{L}_2 = \texttt{undec}(\mathcal{L}_1) \cap \texttt{dec}(\mathcal{L}_2)$

c) if $\mathcal{L}_1 \approx \mathcal{L}_2$ then $\mathcal{L}_1 \ominus \mathcal{L}_2 = \texttt{dec}(\mathcal{L}_1) \cap \texttt{undec}(\mathcal{L}_2) \cup \texttt{undec}(\mathcal{L}_1) \cap \texttt{dec}(\mathcal{L}_2)$

Proof.

a) Follows from the fact that $\texttt{in}(\mathcal{L})$, $\texttt{out}(\mathcal{L})$ and $\texttt{undec}(\mathcal{L})$ partition the domain of any labelling $\mathcal{L}$.
b) and c) are obtained by eliminating the empty sets in a) and replacing $\texttt{in}(\mathcal{L}) \cup \texttt{out}(\mathcal{L})$ by $\texttt{dec}(\mathcal{L})$. □

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

Definition 16 (Hamming set based preference $\geq_{i,\ominus}$). *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.*

Definition 17 (Hamming distance based preference $\geq_{i,|\ominus|}$). *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 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.*

The Hamming set based preference yields a partial order, whereas the Hamming distance based preference yields a total preorder. We now prove two lemmas establishing the relations between less or equally committed labellings and Hamming set/distance based preferences over labellings.

Lemma 2. *Let $\mathcal{L}$, $\mathcal{L}'$ and $\mathcal{L}_i$ be three labellings such that $\mathcal{L} \sqsubseteq \mathcal{L}' \sqsubseteq \mathcal{L}_i$. If $\mathcal{L}_i$ is the most preferred labelling of agent i and his preference is Hamming set or Hamming distance based, then $\mathcal{L}' \geq_{i,\ominus} \mathcal{L}$ and $\mathcal{L}' \geq_{i,|\ominus|} \mathcal{L}$ respectively.*

Proof. From $\mathcal{L} \sqsubseteq \mathcal{L}'$, we have that $\texttt{dec}(\mathcal{L}) \subseteq \texttt{dec}(\mathcal{L}')$, which is equivalent to $\texttt{undec}(\mathcal{L}') \subseteq \texttt{undec}(\mathcal{L})$ because undec is the complement of dec. From this it follows that $\texttt{undec}(\mathcal{L}') \cap \texttt{dec}(\mathcal{L}_i) \subseteq \texttt{undec}(\mathcal{L}) \cap \texttt{dec}(\mathcal{L}_i)$. Since $\mathcal{L} \sqsubseteq \mathcal{L}_i$ and $\mathcal{L}' \sqsubseteq \mathcal{L}_i$ (by assumption and transitivity of $\sqsubseteq$), we can use Lemma 1b to obtain $\mathcal{L}' \ominus \mathcal{L}_i \subseteq \mathcal{L} \ominus \mathcal{L}_i$. By definition we have that $\mathcal{L}' \geq_{i,\ominus} \mathcal{L}$ and $\mathcal{L}' \geq_{i,|\ominus|} \mathcal{L}$. □

Lemma 3. *Let $\mathcal{L}$, $\mathcal{L}'$ and $\mathcal{L}_i$ be three labellings and let $\mathcal{L} \sqsubseteq \mathcal{L}_i$. If $\mathcal{L}_i$ is the most preferred labelling of agent i, his preference is Hamming set based and $\mathcal{L}' \geq_{i,\ominus} \mathcal{L}$, then $\mathcal{L} \sqsubseteq \mathcal{L}'$.*

Proof. $\mathcal{L}' \geq_{i,\ominus} \mathcal{L}$ implies $\mathcal{L}' \ominus \mathcal{L}_i \subseteq \mathcal{L} \ominus \mathcal{L}_i$ which implies $\mathcal{L}(A) = \mathcal{L}_i(A) \Rightarrow \mathcal{L}'(A) = \mathcal{L}_i(A)$ for any argument A (i). $\mathcal{L} \sqsubseteq \mathcal{L}_i$ implies $\mathcal{L}(A) = \mathcal{L}_i(A)$ for any $A \in \texttt{dec}(\mathcal{L})$ (ii). From (i) and (ii) it follows that $\mathcal{L}(A) = \mathcal{L}'(A)$ for any $A \in \texttt{dec}(\mathcal{L})$. Hence $\mathcal{L} \sqsubseteq \mathcal{L}'$. □

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.

4 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. In order to address the question of whether the sceptical and the credulous aggregation operators are Pareto optimal, we first need to define when a labelling Pareto dominates another labelling.

Definition 18 (Pareto dominance). *Let $N = 1, \ldots, n$ be a group of agents with preferences $\geq_i, i \in N$. $\mathcal{L}$* Pareto dominates *$\mathcal{L}'$ if $\forall i \in N$, $\mathcal{L} \geq_i \mathcal{L}'$ 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 19 (Pareto optimality). *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. In particular, candidates for dominance are admissible and less or equally committed labellings in the case of the sceptical operator, and compatible labellings in the case of the credulous operator.

Theorem 2. *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.*

Proof. Let P be a profile of admissible labellings, $\mathcal{L}_{SO} = so_{AF}(P)$ and $\mathcal{L}_X$ some admissible labelling with the property $\forall i \in N, \mathcal{L}_X \sqsubseteq \mathcal{L}_i$. From Theorem 1 we know that $\mathcal{L}_{SO}$ is the biggest admissible labelling with such property, so $\mathcal{L}_X \sqsubseteq \mathcal{L}_{SO}$. So $\forall i \in N, \mathcal{L}_X \sqsubseteq \mathcal{L}_{SO} \sqsubseteq \mathcal{L}_i$. From Lemma 2 we have $\mathcal{L}_{SO} \geq_i \mathcal{L}_X$ for any i. So no agent strictly prefers $\mathcal{L}_X$ and hence there is no labelling that dominates $\mathcal{L}_{SO}$. □

Theorem 3. *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.*

Proof. In the proof of Theorem 2 Hamming set may be replaced by Hamming distance because it is only used in Lemma 2, which works for Hamming distance as well. □

Theorem 4. *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.*

Proof. Let P be a profile of admissible labellings, $\mathcal{L}_{CO} = co_{AF}(P)$, $\mathcal{L}_{CIO} = cio_{AF}(P)$. Assume by contradiction that there exists some admissible labelling $\mathcal{L}_X$ with the property $\forall i \in N, \mathcal{L}_X \approx \mathcal{L}_i$ that dominates $\mathcal{L}_{CO}$.

First notice that compatibility ensures that there are no in/out conflicts between $\mathcal{L}_X$ and $\mathcal{L}_{CO}$. If there is a conflict between agents' labellings on some argument, then both $\mathcal{L}_X$ and $\mathcal{L}_{CO}$ need to label it undec. If there exists an agent whose labelling decides on some argument and other agents' labellings agree or retain from decision, $\mathcal{L}_{CO}$ and $\mathcal{L}_X$ also agree or retain from decision. If all agents retain from decision on some argument, $\mathcal{L}_{CO}$ by definition also retains, and $\mathcal{L}_X$ may label freely. Let us take $A \in \texttt{dec}(\mathcal{L}_X)$. Then, there needs to be an agent with a labelling that agrees on A. Otherwise all agents' labellings would be undecided on such argument and, according to definition, $\mathcal{L}_{CO}$ would not decide either. But then all agents' labellings will agree on such argument with $\mathcal{L}_{CO}$ and disagree with $\mathcal{L}_X$, so no agent will strongly prefer $\mathcal{L}_X$, which contradicts with domination. So there exists at least one agent whose labelling agrees with $\mathcal{L}_X$ on A. Other agents' labellings also need to agree on A or label it undec because of the compatibility of $\mathcal{L}_X$. Then by definition $\mathcal{L}_{CIO}(A) = \mathcal{L}_X(A)$. This holds for any argument $A \in dec(\mathcal{L}_X)$, so we have $\mathcal{L}_X \sqsubseteq \mathcal{L}_{CIO}$. But $\mathcal{L}_X$ is

admissible and, by Theorem 1, $\mathcal{L}_{CO}$ is the biggest admissible labelling less or equally committed as $\mathcal{L}_{CIO}$. So we have $\mathcal{L}_X \sqsubseteq \mathcal{L}_{CO} \sqsubseteq \mathcal{L}_{CIO}$. $\mathcal{L}_X$ must be different from $\mathcal{L}_{CO}$ to dominate it. Let A be an argument on which these labellings differ. From the previous it follows that $A \in \texttt{undec}(\mathcal{L}_X)$ and $A \in \texttt{dec}(\mathcal{L}_{CO})$. $\mathcal{L}_{CO}$ decides on an argument only if there exists an agent that decides on such argument. But then this agent will agree on A with $\mathcal{L}_{CO}$ and disagree with $\mathcal{L}_X$, so it will not prefer $\mathcal{L}_X$. This is in contradiction with dominance. Hence, such dominating labelling cannot exist. □

Observation 1. *The credulous aggregation operator is not Pareto optimal when the preferences are Hamming distance based. In Fig. 2 both labellings $\mathcal{L}_{CO}$ and $\mathcal{L}_X$ are compatible with both $\mathcal{L}_1$ and $\mathcal{L}_2$, but $\mathcal{L}_X$ is closer when applying Hamming distance. $\mathcal{L}_1 \ominus \mathcal{L}_{CO} = \mathcal{L}_2 \ominus \mathcal{L}_{CO} = \{A, B, E, F, G\}$, so Hamming distance is 5, whereas $\mathcal{L}_1 \ominus \mathcal{L}_X = \mathcal{L}_2 \ominus \mathcal{L}_X = \{A, B, C, D\}$, so Hamming distance is 4.*

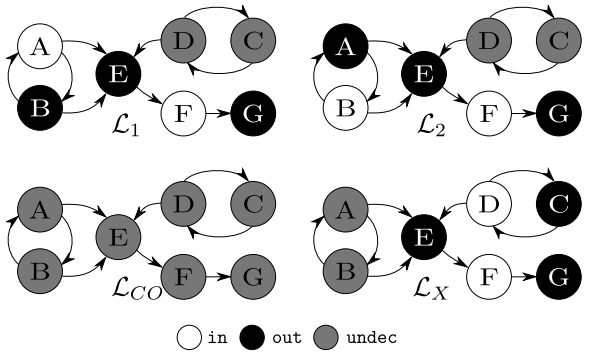

Figure 2: The credulous aggregation operator is not Pareto optimal under Hamming distance based preferences.

We summarise our results in Table 1. We are now ready to address the second research question of the paper, i.e., whether the credulous and sceptical operators are manipulable.

	Sceptical Operator	Credulous Operator
Hamming set	Yes (Theorem 2)	Yes (Theorem 4)
Hamming dist.	Yes (Theorem 3)	No (Observation 1)

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

5 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, an important question to address when dealing with aggregation procedures is to study whether they are strategy-proof (i.e. non-manipulable).

In order to talk about manipulability, we first need to denote a profile in which a labelling has been changed. We recall that by *profile* we refer to a set of individual labellings $\{\mathcal{L}_1, \ldots, \mathcal{L}_n\}$. 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 20 (Strategic lie). *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 21 (Strategy-proof operator). *An aggregation operator O is* strategy-proof *if strategic lies are not possible.*

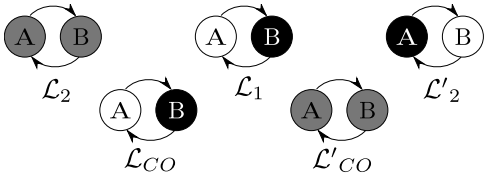

Figure 3: The credulous operator is not strategy-proof.

Observation 2. *The credulous aggregation operator is not strategy-proof. In Fig.3 the agent with labelling $\mathcal{L}_2$ can insincerely report $\mathcal{L}'_2$ to obtain his preferred labelling. This makes an agent with labelling $\mathcal{L}_1$ worse off. The example is valid for both Hamming set and Hamming distance based preferences.*

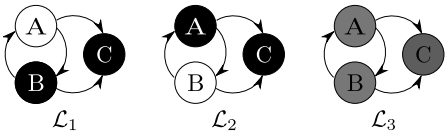

Figure 4: The sceptical operator is not strategy-proof.

Observation 3. *The sceptical aggregation operator is not strategy-proof. Consider the three labellings in Fig. 4. Labelling $\mathcal{L}_1$ of agent 1 when aggregated with $\mathcal{L}_2$ gives labelling $\mathcal{L}_3$, which differs on all three arguments. But, when the agent strategically lies and reports labelling $\mathcal{L}_2$ instead, the result of the aggregation is the same labelling $\mathcal{L}_2$, which differs only on two arguments $\{A, B\}$. The example is valid for both Hamming set and Hamming distance based preferences.*

Surprisingly, however, this lie does not harm the other agent. On the contrary, it improves the social outcome for both the agents. In order to study this kind of situation, we now introduce the distinction between malicious and benevolent lies.

Definition 22 (Malicious lie). *Let O be some aggregation operator and P a profile. We say that a strategic lie $\mathcal{L}'_k$* is malicious *iff, for some agent $j \neq k$, $O(P) >_j O(P_{\mathcal{L}_k/\mathcal{L}'_k})$.*

Definition 23 (Benevolent lie). *Let O be some aggregation operator and P a profile. We say that a strategic lie $\mathcal{L}'_k$* is benevolent *iff, for any agent i $O(P_{\mathcal{L}_k/\mathcal{L}'_k}) \geq_i O(P)$ and there exists an agent $j \neq k$, $O(P_{\mathcal{L}_k/\mathcal{L}'_k}) >_j O(P)$.*

Theorem 5. *Consider the sceptical aggregation operator and Hamming set based preferences. For any agent, his strategic lies are benevolent.*

Proof. Let P be a profile, and $\mathcal{L}'_k$ a strategic lie of agent k. Denote $\mathcal{L}_{SO} = so_{AF}(P)$ and $\mathcal{L}'_{SO} = so_{AF}(P_{\mathcal{L}_k/\mathcal{L}'_k})$. Agent k's preference is $\mathcal{L}'_{SO} >_k \mathcal{L}_{SO}$ (i). We will show that for any agent $i \neq k$, we have $\mathcal{L}'_{SO} >_i \mathcal{L}_{SO}$. Since the sceptical aggregation operator produces social outcomes that are less or equally committed to all the individual labellings, we have that $\mathcal{L}'_{SO} \sqsubseteq \mathcal{L}_i$ for all $i \neq k$ (ii). Similarly, we have $\mathcal{L}_{SO} \sqsubseteq$

$\mathcal{L}_k$ (iii). From (i) and (iii), by Lemma 3, we have that $\mathcal{L}_{SO} \sqsubseteq \mathcal{L}'_{SO}$ (iv). From (iv) and (ii) we have $\mathcal{L}_{SO} \sqsubseteq \mathcal{L}'_{SO} \sqsubseteq \mathcal{L}_i$ for all $i \neq k$. Finally, we can apply Lemma 2 to obtain $\mathcal{L}'_{SO} \geq_i \mathcal{L}_{SO}$ for all $i \neq k$ (v). We showed that lie is not malicious, now we show that it is benevolent. (iii) implies $\texttt{undec}(\mathcal{L}_k) \subseteq \texttt{undec}(\mathcal{L}_{SO})$ (vi). (i) and (vi) implies $\exists A \in \texttt{dec}(\mathcal{L}_k) : A \in \texttt{undec}(\mathcal{L}_{SO}) \wedge A \in \texttt{dec}(\mathcal{L}'_{SO})$ (vii). From (vii), (ii) and (v) $\mathcal{L}'_{SO} >_i \mathcal{L}_{SO}$ for $i \neq k$. □

Theorem 6. *Consider the sceptical aggregation operator and Hamming distance based preferences. For any agent, his strategic lies are benevolent.*

Proof. Let P be a profile, and $\mathcal{L}'_k$ a strategic lie of agent k whose most preferred labelling is $\mathcal{L}_k$. Denote $\mathcal{L}_{SO} = so_{AF}(P)$ and $\mathcal{L}'_{SO} = so_{AF}(P_{\mathcal{L}_k/\mathcal{L}'_k})$. We will show that, if $\mathcal{L}'_{SO}$ is strictly preferred to $\mathcal{L}_{SO}$ by agent k, then it is also strictly preferred by any other agent. Without loss of generality we can take agent $j, j \neq k$,whose most preferred labelling is $\mathcal{L}_j$. Let us partition the arguments into the following disjoint groups: $A = \texttt{dec}(\mathcal{L}_{SO}) \backslash \texttt{dec}(\mathcal{L}'_{SO})$ (decided arguments that became undecided), $B = \texttt{dec}(\mathcal{L}'_{SO}) \setminus \texttt{dec}(\mathcal{L}_{SO})$ (undecided arguments that became decided), $C = \texttt{dec}(\mathcal{L}'_{SO}) \cap \texttt{dec}(\mathcal{L}_{SO})$ (arguments decided in both labellings), $D = \texttt{undec}(\mathcal{L}'_{SO}) \cap \texttt{undec}(\mathcal{L}_{SO})$ (arguments undecided in both labellings). Labellings $\mathcal{L}_{SO}$ and $\mathcal{L}'_{SO}$ agree on the arguments in D (which are labeled `undec`) and C, whose arguments are labeled `in` or `out`. On the arguments in C there are no $\texttt{in} - \texttt{out}$ conflicts between $\mathcal{L}_{SO}$ and $\mathcal{L}'_{SO}$ as the sceptical aggregation operator guarantees social outcomes less or equally committed than $\mathcal{L}_j$. Therefore, only arguments from A and B have an impact on Hamming distance. Both labellings $\mathcal{L}_k$ and $\mathcal{L}_j$ agree with $\mathcal{L}_{SO}$ on the arguments in A because $\mathcal{L}_{SO}$ decides on those arguments and is less or equally committed than both labellings. $\mathcal{L}'_{SO}$ remains undecided on the arguments in A so both labellings $\mathcal{L}_k$ and $\mathcal{L}_j$ disagrees with $\mathcal{L}'_{SO}$ on A. $\mathcal{L}'_{SO}$ is less or equally committed than $\mathcal{L}_j$ so, as above, we obtain that on the arguments in B, $\mathcal{L}_j$ agrees with $\mathcal{L}'_{SO}$ and disagrees with $\mathcal{L}_{SO}$. On the contrary, $\mathcal{L}'_{SO}$ does not have to be less or equally committed than $\mathcal{L}_k$ and so, for agent k, some of the arguments from B increase the distance and some of them decrease it. If agent k prefers $\mathcal{L}'_{SO}$ to $\mathcal{L}_{SO}$, then the number of the arguments decreasing the distance must be greater than the number of those increasing by more than $|A|$. But for agent j all the arguments from B are decreasing the distance, as $\mathcal{L}_j$ agrees with $\mathcal{L}'_{SO}$ on the whole B. So, if agent k gains by switching to labelling $\mathcal{L}'_{SO}$, agent j needs to gain by at least the same. □

We summarise our results in Table 2.

	Sceptical	Credulous
Hamming set	No (Obs. 3) but benev. (Th. 5)	No, and not benev. (Obs. 2)
Hamming dist.	No (Obs. 3) but benev. (Th. 6)	No, and not benev. (Obs. 2)

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

6 Conclusion and related work

The study of aggregation problems in abstract argumentation is recent. [Coste-Marquis *et al.*, 2007] present an approach to merge Dung's argumentation frameworks. The argumentation frameworks to be merged may be different, that is agents may ignore arguments put forward by other agents.

Given an argumentation framework, [Rahwan and Tohmé, 2010] 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 [Rahwan and Larson, 2008], 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`.

References

[Brams *et al.*, 2007] S Brams, M Kilgour, and R Sanver. A minimax procedure for electing committees. *Public Choice*, 132(3-4):401–420, 2007.

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

[Caminada, 2006] M.W.A. Caminada. On the issue of reinstatement in argumentation. In *JELIA 2006*, pages 111–123. Springer, 2006. LNAI 4160.

[Coste-Marquis *et al.*, 2007] 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.

[Dung, 1995] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

[Konieczny and Perez, 1998] S. Konieczny and R. Pino Perez. On the logic of merging. In *Proceedings of KR'98*, pages 488–498, 1998.

[Pigozzi, 2006] G. Pigozzi. Belief merging and the discursive dilemma: an argument-based account to paradoxes of judgment aggregation. *Synthese*, 152(2):285–298, 2006.

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

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