Argument & Computation 8 (2017) 1–34
DOI 10.3233/AAC-160014
IOS Press

Collective argumentation: A survey of aggregation issues around argumentation frameworks

Gustavo Bodanza [a,*], Fernando Tohmé [b] and Marcelo Auday [a]

[a] *Departamento de Humanidades, Universidad Nacional del Sur and IIESS-CONICET, Argentina*
[b] *Departamento de Economía, Universidad Nacional del Sur and INMABB-CONICET, Argentina*

Abstract. Dung's argumentation frameworks have been applied for over twenty years to the analysis of argument justification. This representation focuses on arguments and the attacks among them, abstracting away from other features like the internal structure of arguments, the nature of utterers, the specifics of the attack relation, etc. The model is highly attractive because it reduces most of the complexities involved in argumentation processes. It can be applied to different settings, like the argument evaluation of an individual agent or the case of dialectic disputes between two agents (pro and con), or even in multi-agent collective argumentation. The latter case involves agents with possibly different arguments and/or opinions on how to evaluate them, leading to the possibility of considering multiple sets of arguments and attack relations. Two basic questions can be asked here, namely 'what to aggregate' and 'how to aggregate'. The former concerns what kinds of entities do the agents intend to choose (arguments, attacks, assessments, etc.), while the second one focuses on which aggregation mechanisms yield rational choices (voting on arguments, merging procedures to obtain a common argumentation framework, deliberation processes, etc.). In particular, the question about the rationality of a collective argument choice relates this topic to Social Choice and Judgment Aggregation theories, while its associated strategic issues relate it to Game Theory. The research efforts on the disparate problems elicited by collective argumentation have generated a considerable corpus of literature that deserves an orderly evaluation. This survey is intended as a contribution to that end.

Keywords: Argumentation frameworks, collective argumentation, merging argumentation systems, multi-agent systems, social choice, judgment aggregation, deliberation, game theory

1. Introduction

The study of argumentation processes has a long history, covering many different issues, with a scope that ranges from Logics to Rhetoric. It received a new impetus thanks to the problems that arose in the study of Knowledge Representation systems in Artificial Intelligence. One of the main contributions in that field was the characterization of *abstract argumentation frameworks* by Dung [30].

The main idea of Dung's model is that agents can present arguments pro and con other arguments, and the sets of arguments that are "coherent" in some strong sense can be seen as possible outcomes of argumentation processes. In such sets, the arguments do not counter each other and, furthermore, any attacked argument can be "defended" by other arguments in the set. The formalization of this idea involves a class of arguments (without any specific internal structure) and an attack relation among them. The *extensions* of this framework, which formalize the notion of coherence, are defined in terms of the attacks among arguments.

[*] Corresponding author. E-mail: bodanza@gmail.com.

1946-2166/17/$35.00 © 2017 – IOS Press and the authors.

The literature that grew out of Dung's original contribution amounts to several thousands of papers. A fraction of this large corpus is devoted to the analysis of problems of *collective argumentation*, i.e. argumentation in which several frameworks are involved. In a broad sense the aim is to find the acceptable outcomes when agents with either different sets of arguments and/or different attack relations interact in an argumentation process.

The research efforts on collective argumentation problems have also generated a considerable corpus of literature deserving an orderly evaluation. This paper is intended as a contribution in that direction. We distinguish several streaks in the literature, each characterized by a different modeling feature. Even so, some important topics related to the field are left out of our study, in particular the contributions based on argumentation models different from Dung's and the large literature relating negotiation with argumentation.

1.1. Organization of the paper

In Section 2 we briefly describe Dung's argumentation frameworks, the encompassing formalism of our study. In Section 3 we overview the motivations leading to the study of argument aggregation. In Section 4 we analyze the contributions in the literature of collective argumentation, distinguishing between those that are based on the interaction of different sets of arguments and those on different attack relations. Section 5 discusses the application of ideas and methods from Game Theory and Social Choice Theory to collective argumentation. Section 6 analyzes the literature on dynamic argumentation. Section 7 analyzes the literature in which the conflicts of collective argumentation are solved by means of the use of numerical weights attached to components of the aggregate system. Some prominent works are detailed in Section 8. Section 9 covers the works that do not precisely fit in the previous taxonomy. We conclude in Section 10 with some considerations about possible future work.

2. Dung's argumentation frameworks

The highly abstract character and the simplicity of Dung's argumentation frameworks are surely some of the main reasons for the adoption of this model in studies on argument aggregation. Arguments are primitives of the model without any assumed internal structure. The attack relation among them, a binary relation without any further restriction, is also a primitive of the model, independent of the possible reasons for the attacks.

Definition 1. An *argumentation framework* is a pair $AF = \langle AR, \rightharpoonup \rangle$, where AR is a set of abstract entities called 'arguments' and $\rightharpoonup \subseteq AR \times AR$ denotes an attack relation among arguments.

Arguments interact through the attack relation. To determine which arguments survive, different characterizations of the notion of defense yield disparate sets of supported arguments, the *extensions* of AF. These extensions are seen as the semantics of the argumentation framework, i.e. the classes of arguments that can be deemed as the outcomes of the whole process of argumentation. Dung introduces the notions of *preferred*, *stable*, *complete*, and *grounded* extensions.

Definition 2. Given an argumentation framework AF, an argument a is said *acceptable* w.r.t. a subset S of arguments of AR, if for every argument b such that $b \rightharpoonup a$, there exists some argument $c \in S$ such that $c \rightharpoonup b$. A set of arguments S is said *admissible* if each $a \in S$ is acceptable w.r.t. S and is conflict-free, i.e., the attack relation does not hold for any pair of arguments belonging to S. Then we say that S is:

- a *preferred extension* if it is any maximally (w.r.t. set inclusion) admissible set of arguments of AF;
- a *complete extension* of AF if it is any conflict-free subset of arguments which is a fixed point of $\Phi(\cdot)$, where $\Phi(S) = \{a : a \text{ is acceptable w.r.t. } S\}$;
- a *grounded extension* if it is the least (w.r.t. $\subseteq$) complete extension;
- a *stable extension* if it is a conflict-free set S of arguments which attacks every argument not belonging to S.

An important result, of relevance in several works on collective argumentation, is that all these extension semantics coincide in argumentation frameworks which are *well-founded*, that is, in which there does not exist a infinite sequence $a_1, a_2, \ldots, a_n, \ldots$ such that a_{i+1} attacks a_i. For argumentation frameworks with finite sets of arguments this means that the attack relation has no cycles.

For a discussion on extension semantics for argumentation frameworks and alternative proposals to Dung's semantics we refer the reader to Baroni, Caminada, and Giacomin [10].

Another way of defining the status of arguments in argumentation frameworks is given by a labelling semantics. A labelling is a representation of an argumentation framework in which each argument is labeled in, out or undec according to its interactions with the other arguments in a framework [10, 19,44,73]. Let us consider in some detail Caminada's definitions [19].

Let $AF = \langle AR, \rightharpoonup \rangle$ be an argumentation framework. A *labelling* of AF is a total function $\mathcal{L} : AR \longrightarrow \{\mathtt{in}, \mathtt{out}, \mathtt{undec}\}$. A labelling $\mathcal{L}$ is *admissible* iff for every argument $a \in AR$, $\mathcal{L}(a) = \mathtt{in}$ iff $\mathcal{L}(b) = \mathtt{out}$ for every $b \in AR$ such that $b \rightharpoonup a$, and $\mathcal{L}(a) = \mathtt{out}$ iff $\mathcal{L}(b) = \mathtt{in}$ for some $b \in AR$ such that $b \rightharpoonup a$. A labelling $\mathcal{L}$ is *complete* iff it is admissible and for every argument $a \in \mathcal{L}$, $\mathcal{L}(a) = \mathtt{undec}$ iff $\mathcal{L}(b) \neq \mathtt{in}$ for every $b \in AR$ such that $b \rightharpoonup a$, and $\mathcal{L}(b) = \mathtt{undec}$ for some $b \in AR$ such that $b \rightharpoonup a$. The sets of all the arguments labeled in, out and undec by a labelling $\mathcal{L}$ will be denoted by $\mathtt{in}(\mathcal{L})$, $\mathtt{out}(\mathcal{L})$ and $\mathtt{undec}(\mathcal{L})$, respectively.

Caminada [19] states formal correspondences among labelling and extension semantics, in the sense that the set of all the arguments labeled in in a given labelling is an extension of a given semantics. More precisely,

- complete labellings correspond to complete extensions,
- complete labellings with empty undec correspond to stable extensions,
- complete labellings with maximal in correspond to preferred extensions,
- complete labellings with maximal undec correspond to grounded extensions.

3. Motivations for the study of argument aggregation

The aggregation of opinions is clearly the most general and inspiring motivation for the study of argument aggregation. This motivation is shared in part by Social Choice Theory (SCT) [4], which is devoted to the study of preference aggregation, and by Judgment Aggregation Theory (JAT) [43,54–56], which studies the aggregation of opinions in the form of sentences. In this regard, we can say that collective argumentation is concerned with the explicit or implicit aggregation of individual preferences among arguments in order to find collective opinions based on collectively supported reasons. Deliberative democracy is clearly reflected in this characterization [62], but models of argument aggregation can potentially be used for a wider range of applications covering, for instance, collective intelligence [8] and prediction markets [60].

This general problem can be decomposed in several subsidiary problems, among which the one prevailing in the literature is that of which arguments should be collectively selected. We can also find some models not of collective selection but of collective assessment of arguments. Finally, other models are motivated by information exchange or knowledge expansion rather than by selection or assessment problems.

Taking into account those differences we identified what we consider are the main motivations behind collective argumentation. The categories are not exclusive but inclusive, meaning that some works can be in line with more than one of these motivations. Consequently, different methods and techniques may appear combined in disparate ways.

Aggregation mechanisms. Many of the works in the literature are devoted to the design of mechanisms for collective argument selection. To this end, every approach responds to two questions: what to aggregate (e.g. arguments, attack criteria, argumentation frameworks) and how to aggregate (e.g. voting mechanisms, merging procedures, deliberation processes). This is discussed in Section 4.

Rationality properties. Another important research line is devoted to the study of the properties required to deem as rational the outcomes of aggregation mechanisms. Some of them are familiar in fields as SCT or JAT, where impossibilities of either jointly satisfying a set of desirable properties (a typical SCT problem) or of getting collectively acceptable choices under collectively acceptable reasons (a typical JAT problem) have been extensively analyzed. Affinities with Game Theory (GT) can also be found in this literature, particularly in the study of the problem of finding strategy-proof aggregation mechanisms. All these issues are treated in Section 5.

Dynamic argumentation. Argumentation can naturally be understood as a deliberative process in which the status of arguments changes along with the system's evolution. Some works seek to characterize the rationality of the changes, trying an AGM-style axiomatic approach. Other contributions present particular protocols or algorithms capturing deliberative processes. Moreover, there also exist a few papers that are not so much motivated by the collective choice of arguments but by the way agents exchange arguments to get knowledge expansion. We comment on these issues in Sections 6 and 9.

Social argument assessment. Some authors have been concerned with finding a metrics for the quantitative expression of social assessments of arguments, as a way of finding to what extent an argument is collectively supported. This issue is treated in Section 7.

Table 1 summarizes the above motivations including corresponding trending concerns and usually employed techniques.

Although we cannot discard negotiation as a significant motivation for the study of argument aggregation, we decided to leave it out of this review since it has generated a specific, stand alone, literature. We refer the interested reader to [29].

Table 1

Current motivations in the study of argument aggregation

MOTIVATION	Trending concerns	Some employed techniques
Aggregation mechanisms	argument-wise vs. framework-wise approaches	voting mechanisms, procedures, etc.
Rationality properties	analogous to SCT, JAT and GT	particular techniques from those fields
Dynamic argumentation	characterizing changes in the status of args. through deliberative processes	debate models, deliberative proc., AGM-style postulates, modal logics
Social argument assessment	quantifying agreement and disagreement	using numerical values (weights, strengths, distances, etc.)

4. Aggregation mechanisms: Argument-wise and framework-wise methods

Collective argumentation is in many respects a very complex issue. In general, an obvious way of dealing with complexity is to study a few variables assuming (*ceteris paribus*) that the others remain fixed. But even so, researchers are usually confronted with a plethora of possible choices. In our case, just consider the ontology of an argumentation framework. We have arguments and an attack relation, which can be complemented with concepts like preference relations, extensions, labellings, values, weights, etc., and this just in systems that are kept abstract to some extent. In less abstract systems, we could also find: a) components of arguments, like rules, premises, conclusions; b) attack relations decomposed into conflict and preference relations; and c) types of arguments according to the values they promote, etc. Furthermore, in multi-agent systems – where the problem of collective argumentation was originally introduced [57] – we can find several interacting agents, pondering the arguments according to different goals, plans, interests, desires, biases, etc. This profuse ontology poses a first problem for collective argumentation, namely the problem of *what to aggregate*, being some of the alternatives: arguments, preferences on arguments, defeat criteria, labellings, extensions, etc. Moreover, a further problem is given by the question of *how to aggregate*, which leads to analyze voting mechanisms, dynamical selection processes, etc.

In turn, each of these problems involves a variety of different intuitions that have been approached in disparate ways. One basic approach seeks to find a set of collectively justified arguments in a similar way as judgments are aggregated in the field of judgment aggregation. That is, agents can just vote on arguments to decide which of them are collectively accepted. In abstract argumentation frameworks this can be modeled in basically two ways: aggregating extensions or aggregating labellings.[1] We can call this the *argument-wise* aggregation method. Under this view, the questions of what and how to aggregate are answered as follows: individually supported arguments must be aggregated by some voting mechanism.

A second view amounts to find a collective criterion for determining, for instance, whether a preference among arguments is relevant for a given collective decision problem, or whether an attack occurs between arguments, or if an attack can be overlooked since it comes from an argument which promotes a pointless value, and so on. This means that this view rests upon the intuition that collective decisions should be founded on agreements on the way in which arguments interact through attacks. This amounts to finding a common argumentation framework up from a profile[2] of argumentation frameworks. The aggregation comes from merging all the argumentation frameworks into one. Accordingly, we can call this the *framework-wise* aggregation method. Under this view, the questions of what and how to aggregate are then answered in this way: individually supported criteria (represented by different attack relations or even entire argumentation frameworks) will be aggregated by means of some "merging" method.

The approaches to the what and how to aggregate problems in the current literature can be roughly classified along these two views, the argument-wise and the framework-wise one. But we have to emphasize that we do not consider them categories in a strict taxonomy. This is because in some works the particular techniques employed are combined in such a way that it becomes impossible to say which view prevails. But in general this distinction remains useful for a rough identification of underlying intuitions.

[1] Since each Dung's semantics is in correspondence with a kind of restriction over labellings [19], one can expect that both methods lead to the same collectively supported arguments.

[2] We use here the term 'profile' as in SCT: a vector of items $(x_1, \ldots, x_n)$, commonly preference relations, each x_i corresponding to an individual $i \in N$, where N is the set of individuals belonging to the society. So, a social profile describes the individual preferences in a state of the society.

Let us consider a very simple example illustrating both approaches:

Example 1. Let $N = \{1, 2, 3\}$ represent a group of three agents deciding which among three arguments, a, b, and c, are collectively acceptable.[3] Assume that each agent has a subjective evaluation of the interaction among those arguments, leading to three different individual argumentation frameworks:

$$AF_1 = \langle \{a, b, c\}, \{(a, b), (b, c)\} \rangle,$$
$$AF_2 = \langle \{a, b, c\}, \{(a, b)\} \rangle,$$
$$AF_3 = \langle \{a, b, c\}, \{(b, c)\} \rangle.$$

Then an argument-wise approach can be implemented by, first, obtaining the extensions of each framework and, second, aggregating those extensions to obtain the collectively supported ones. For instance, using grounded semantics we will obtain the extensions $\{a, c\}_1$, $\{a, c\}_2$, and $\{a, b\}_3$ of AF_1, AF_2, and AF_3, respectively. Then a voting mechanism can be applied. By majority voting, for instance, we can get the "collective" extension $\{a, c\}$ (in this case, the same result is obtained by counting either the individual extensions or the individually accepted arguments). On the other hand, a framework-wise approach can be implemented by, first, obtaining a "collective" argumentation framework and, second, obtaining the extensions of that framework. Having a common set of arguments, the voting can just be applied to obtain a common attack relation from the individual ones. By majority voting on the pairs proposed by the agents we can get the attack relation $\{(a, b), (b, c)\}$. Thus, the collectively accepted arguments are those corresponding to the extensions of the "collective" argumentation framework $\langle \{a, b, c\}, \{(a, b), (b, c)\} \rangle$. Considering grounded semantics, for instance, we get the extension $\{a, c\}$.

Questions related to which approach, the argument-wise or the framework-wise, is more rational, fair, accurate, etc. are subject of deep philosophical analyses yet to be carried out. Only a few discussions on this topic can be found in the current literature. Coste-Marquis, Devred, Lagasquie-Schiex, Konieczny, and Marquis [25,57], for instance, refer to argument-wise methods as "naive" and favor framework-wise procedures. The simple idea of voting on selected extensions is criticized, showing an anomalous case in which arguments that are not accepted by neither agent become collectively selected. The reason for rejecting voting on extensions is that this method involves a loss of information, since once obtained the extensions the underlying attack relations are disregarded. On the other hand, this voting method – the authors argue – could only be sensible if all the agents start with the same set of arguments. In fact, some works based on the aggregation of labellings (see Section 4.2.1) study argument-wise methods finding also drawbacks. Rahwan and Larson [65], for instance, show that if some arguments are known by some but not all the agents, the ones who know can manipulate the outcome by choosing which arguments to reveal and which to hide (more in Section 5). On the other hand, Delobelle, Haret, Konieczny, Mailly, Rossit, and Woltran [27] stand for an extension-based (argument-wise) approach when several groups of experts debate on a common topic. These authors argue: "Each expert group delivers a set of arguments to be jointly accepted. In order to combine these proposals, the actual structure of the debate of each group is not central; rather, it is the sets of extensions we need to combine".

No further discussions on the virtues and shortcomings of both procedures can be found in the literature. In the next subsections we comment on the particular approaches that can be classified as being either framework-wise or argument-wise.

[3] Cases where the agents consider different arguments introduce further concerns, as we will see later.

4.1. The framework-wise approach

The pioneering work of Coste-Marquis et al. [25,57] on merging argumentation systems was motivated by the problem of deriving sensible information up from a collection of argumentation systems belonging to different agents. These authors focus on scenarios in which some agents are able to consider arguments not known by other agents and disagree on the attack relation. As pointed out above, these works discard argument-wise methods and propose instead a three-step process. In the first one, each attack relation is consensually expanded to become a partial system over the entire set of arguments considered by the group of agents. The second step amounts to merge the expanded systems, generating a class of argumentation systems that are at the shortest "distance" of the ones in the profile. The final step consists in selecting the acceptable arguments from the argumentation systems determined in the previous step (this work is detailed in Section 8.1).

Consensual expansions are also considered by Bromuri and Morge [18]. They present a multi-party argumentation game using event calculus [50]. The authors describe their work as an "individual-based approach where the cross-fertilization of argumentations emerge from the interactions between the agents". In fact, the authors are not aimed at finding the aggregation of a profile of individual argumentation frameworks into a collectively supported one. They just define a way by which each agent could obtain a *partial argumentation framework* (PAF)[4] on the basis of its own initial argumentation framework through a consensual expansion. The "gameboard" is a common environment consisting basically of an initial set of arguments and an attack relation, in which the agents can act by adding arguments, and either adding or deleting attacks. At the end of the game, each agent obtains her own expansion by confronting her initial framework with the common PAF obtained in the game (this differs from the approach by Coste-Marquis et al., where the consensual expansions are obtained by comparing each individual PAF with the entire profile). The game has two subgames, the argument game and the attack game. In the argument game, the agents add arguments until no agent makes any further move; in the attack game, every agent either adds or delete attacks until every agent withdraws. When this happens, the whole game ends.

Gabbay [37] deals with several issues involved in the interaction among argumentation frameworks. The merging problem is conceived as a particular instance in *fibered* argumentation systems (see also [36]). Gabbay considers the case in which the merged system $\langle AR, \rightharpoonup \rangle$ is simply such that $AR = \bigcup_{i \in N} AR_i$ and $\rightharpoonup = \bigcup_{i \in N} \rightharpoonup_i$, where AR_i and $\rightharpoonup_i$ are, respectively, the set of arguments and the attack relation of agent $i \in N$. He poses the problem of finding the complete extensions of the fibered system up from the known extensions of the agents' systems (it is not clear whether this involves also an argument-wise procedure). In Gabbay's approach a priority order among agents is defined which can be used in the merging process. Furthermore, it gives more weight to some agents in the case that a voting system is implemented.

The work by Toni and Torroni [72] can be seen as a particular instance of Gabbay's fibered systems. The authors propose a methodology for analyzing, in computational terms, the exchanges produced in social networks, technical fora and e-commerce sites. They observe that most of the existing work considering online systems and argumentation focuses on extracting argumentation frameworks of one form or another manually or semi-automatically from these exchanges, while computational argumentation has focused on the dialectical acceptance of arguments (or claims supported by arguments) with respect to a statically defined argumentation framework. Given that, the authors propose to fill the gap between

[4] Formally defined in Section 8.1.

those two lines of work. To this end, they envisage that the active participants in the exchange make very simple and graphical annotations, cataloguing texts as *comments* or *opinions* and drawing *links* to indicate source, support or objection. The annotations are expected to be dynamically added by users as the exchanges are performed over time in existing online systems. The "bottom-up argumentation" refers to the way the argumentation framework is built from users' opinions, comments and suggested links. The authors choose the Assumption-Based Argumentation model [31] – an instance of Dung's argumentation frameworks – for an automated mapping of the annotations in order to get the automatic computation of their dialectical acceptance. Clearly, though indirectly, the argumentation framework is the result of the aggregation (addition) of different users' annotations. An open problem posed by the authors is whether the model would lead to different results when applied to the independent markings of other groups of users. This may call for a specific aggregation protocol to overcome the possible differences of opinion.

Pedersen and Dyrkolbotn [61] introduce a stepwise merging mechanism to model a deliberative process satisfying *faithfulness*, a condition requiring that every attack in the merged system must be supported by at least one agent. The authors are concerned with the conciliation of opinions in the argumentative process involving several agents. A novelty of the approach is the use of a dynamic modal logic language (previously used to model argumentation dynamics, e.g. Grossi [42]) to express the possible ways in which agents can deliberate in order to reach an agreement on the attacks among arguments.

The work of Airiau, Bonzon, Endriss, Maudet, and Rossit [1] presents an interesting variation. Instead of focusing on how to aggregate a profile of individual argumentation frameworks into one, they ask how a profile can be *rationalized* up from a "master" argumentation framework. The intuition behind this is as follows. Assume each agent i elaborates her own argumentation framework AF_i considering the values that the arguments promote and her subjective assessment of those values, expressed by a preference order $\succeq_i$. Consider now an argumentation framework AF, the "master" argumentation framework. If AF yields AF_i under $\succeq_i$, then, in a sense, AF_i can be explained by AF (and $\succeq_i$). If that can be done for every single individual argumentation framework in the profile, then AF *rationalizes* (provides a rationale for) the profile.

Other works on merging argumentation frameworks are devoted to specific issues, which will be discussed below. The works by Tohmé, Bodanza, and Simari [71], Dunne, Marquis, and Wooldridge [33], and Endriss and Grandi [35] (the latter is actually on the more general topic of graph aggregation) analyze the satisfiability of rationality conditions in consonance with principles drawn from SCT, so they will be discussed in Section 5. Gabbay and Rodrigues [38] try to define the weight or strength of arguments in a merged system according to the weights given by the agents associated to the local systems. Since the approach is numerical, it will be discussed in Section 7.

4.2. *The argument-wise approach*

4.2.1. *Aggregating labellings*

A recent strain in the literature on argument-wise aggregation deals with the aggregation of labellings (Awad, Booth, Tohmé, and Rahwan [9], Rahwan and Larson [66], Rahwan and Tohmé [69], Caminada and Pigozzi [20], Bodanza [11], Booth, Awad, and Rahwan [15]).

Intuitively, the aggregation of labellings consists in finding a labelling of a given argumentation framework that corresponds to ("represents") a given non-empty set of labellings of the argumentation framework. That is, if $\mathcal{L}_{AF}$ is the set of all the possible labellings of a given argumentation framework AF, a *general labellings aggregation operator* is a function F_{AF}: $2^{\mathcal{L}_{AF}} \setminus \{\emptyset\} \longrightarrow \mathcal{L}_{AF}$.

Given an argumentation framework AF and a set of agents $N = \{1, 2, \ldots, n\}$, the evaluation that each individual $i \in N$ makes of the arguments of AF can be modeled by a labelling $\mathcal{L}_i$. The question is, given a profile of "reasonable" labellings, $(\mathcal{L}_1, \ldots, \mathcal{L}_n)$, what individual labelling (not necessarily in the profile) can be deemed a "reasonable" outcome of the aggregation?[5] Here by "reasonable" we mean that it satisfies a set of integrity constraints, including some semantical requirements such as, for example, that the collective labelling should be admissible provided that all the individual labellings are admissible too. Other constraints usually stem from related fields as SCT or JAT (we discuss these conditions in Section 5).

Note the fundamental difference between the aggregation of labellings and the aggregation of frameworks. In the former the agents put labels on a commonly shared objective argumentation framework, and the problem is to find a common labelling representing the profile of individual labellings. In the aggregation of frameworks, instead, each agent builds an individual framework (where both the set of arguments and the attack relation can be introduced at will by the agent) and the problem is to find a common argumentation framework representing the profile of those individual frameworks.

The way in which labellings are defined presents a limitation for their aggregation. If the argumentation framework at stake is well-founded – i.e. the attack relation has no cycles, making all of Dung's semantics identical, yielding a single extension – there exists only one admissible labelling. Thus, the aggregation problem only becomes interesting in non well-founded argumentation frameworks. But the aggregation of labellings may become nontrivial even in well-founded settings if we allow the agents to introduce new arguments and attacks. For instance, Rahwan and Larson [65,66] consider *types* of agents, such that the type θ_i of agent i includes the arguments *available to* i (together with the preferences of i with respect to the possible outcomes of the aggregation). If not all the arguments of the framework are available to i then she will produce labellings that are possibly at odds with the acceptable labellings of the whole framework. In sum, interesting cases are those where either the collective argumentation framework is not well-founded or the agents are not aware of all the arguments at stake.

4.2.2. *Aggregating extensions*

Delobelle et al. [27] is the only work – as far as we know – that introduces specific merging operators based on extensions, though in combination with a framework-wise merging process. These authors study the generation of the argumentation framework resulting from a merging process. Given a semantics σ, a profile of argumentation frameworks and a set $\mathcal{C}$ of candidates (i.e. a class of sets of arguments), a generation operator yields a set of argumentation frameworks such that their σ-extensions are the elements of $\mathcal{C}$. The approach exploits the method presented in Coste-Marquis et al. [25,57] based on the use of a pseudo-distance among argumentation frameworks (see Section 7). Since the latter work also deals with the rationality postulates underlying merging and revision operators, we will go back to this in Sections 5 and 6.

4.3. *The commutation problem*

Beyond the question of which procedure is more suitable, Bodanza and Auday [12] analyze the *commutation* problem: to determine which conditions ensure that the extensions of an aggregate argumentation framework obtained through an argument-wise procedure are the same as the aggregate extensions of the profile of individual argumentation frameworks obtained through a framework-wise procedure. The problem can be illustrated by a slight modification of Example 1.

[5] In some works (e.g. Caminada and Pigozzi [20]) a profile of labellings can be treated just as a set. See Section 8.4.

Example 2. Let $N = \{1, 2, 3\}$ represent a group of three agents, each one represented by an argumentation framework as follows:

$$AF_1 = \langle \{a, b, c\}, \{(a, b), (b, c)\} \rangle,$$
$$AF_2 = \langle \{a, b, c\}, \{(c, b), (b, a)\} \rangle,$$
$$AF_3 = \langle \{a, b, c\}, \{(b, a), (b, c)\} \rangle.$$

Through an argument-wise approach, based on majority voting and considering grounded semantics, we obtain the extensions $\{a, c\}_1$, $\{a, c\}_2$, and $\{b\}_3$ of AF_1, AF_2, and AF_3, respectively, from which the "collective" extension $\{a, c\}$ is obtained. On the other hand, through a framework-wise approach (using also majority voting and grounded semantics) the "collective" argumentation framework $\langle \{a, b, c\}, \{(b, a), (b, c)\} \rangle$ is obtained which yields the extension $\{b\}$.

The authors find some (rather severe) restrictions preserving the equivalence between the mechanisms, for instance: the class of arguments must not have more than three arguments; the number of agents must be odd; the decision is obtained by absolute majority; the individual attack relations are irreflexive and asymmetric; for every pair of arguments a and b, all the agents agree on either, 1) a and b are in conflict (i.e. either $a \rightharpoonup b$ or $b \rightharpoonup a$), or 2) a and b are not in conflict; and the agents unanimously agree on the attack relation on at least one pair of arguments. Auday [6] establishes other rather negative results: commutation is not possible if the aggregation procedure respects *strict unanimity* (i.e. a social outcome is obtained if and only if all the agents agree on it), not even in the most basic settings, i.e. with only two arguments, two agents and acyclic attack relations. This negative result extends to aggregation procedures in which the proportion of agents that have to agree to impose a collective outcome has to reach a certain *quota*. The same is true for *qualified voting* aggregation rules, in which only a selected group of agents has to agree to impose the outcome. On the other hand, Auday [7] finds a positive result for revealed individual attack relations (i.e. subsets of pairs of arguments – the attacks "revealed" by the agents – that yield the same extensions as the original attack relation) under grounded semantics. If the original attack relations are all acyclic, the grounded extension of the union of the corresponding revealed attack relations coincides with the intersection of the grounded extensions of the revealed attack relations.

5. The rationality of argument aggregation: Applying concepts from social choice theory, judgment aggregation theory and game theory

SCT is concerned with the elicitation of a collective preference up from a profile of individual preferences. On the other hand, JAT is devoted to find a collectively acceptable judgment on a given set of propositions (the *agenda*) stemming from a profile of individual judgments. SCT and JAT have clear similarities but also differences. JAT is in some sense analogous to SCT since they are both concerned with aggregating opinions satisfying certain properties. On the other hand, judgment aggregation is different from preference aggregation since judgments may be logically related among them and, hence, the ensuing aggregate set of propositions should satisfy logical properties in addition to social choice-theoretic ones. Argument aggregation, in turn, can be viewed as a special case of JAT. The sentence 'argument a is `in`' can be seen as a judgment about the acceptance status of a; the sentence 'argument a attacks argument *b*' can be seen as a judgment about the acceptance of the attack of *a* on *b*. The aggregation of a set of judgments like these should satisfy some properties in order to guarantee the rationality of the aggregate. For instance, if both sentences above are collectively accepted, then the sentence 'argument *b* is `in`' should not be collectively accepted since the aggregate set of judgments must be consistent. On the other hand, if all the agents approve unanimously a given judgment it is reasonable to demand that this judgment should be included in the aggregate set. The latter is a typical social choice-theoretic property, which indicates that argument aggregation involves rationality properties proper of both JAT and SCT.

5.1. Properties from SCT and JAT studied in argument aggregation

We list some common properties from SCT (cf. Arrow, Sen, and Suzumura [5]) and JAT (cf. List and Puppe [56]) discussed in the literature on argument aggregation. We make an informal presentation in order to make these concepts more intuitive.

From SCT:

- *Unanimity*: if all agents agree on an alternative, the aggregate outcome must also agree with them.
- *Positive responsiveness*: the aggregation function should keep yielding the same outcome if some agent, previously against it, now is in its favor. (Intuitively, adding votes to an already winning alternative does not make it a losing one.)
- *Anonymity*: the outcome of a given profile should be the same for any permutation of the profile, i.e. the outcome depends on the choices of the agents but not on who they are.
- *Pareto optimality*: an outcome is Pareto optimal if no other outcome is at least as preferred for every agent, and strictly preferred for some agent.
- *Independence of irrelevant alternatives*: the relation between a pair of alternatives in the aggregate does not change with the introduction of a third alternative irrelevant to the other two. (Some intuition from political elections may be useful: if the individual preferences over two candidates *a* and *b* remain the same when a third candidate *c* arises, the rank of *a* and *b* should be the same in elections with and without *c*. That is, the resulting relation between *a* and *b* in the result is independent of *c*, which is irrelevant to the other two in the individual preferences.)
- *Non-dictatorship*: there is no agent i_0 such that her choices become the outcome in every possible instance. That is, there is no 'dictator' among the agents.

From JAT:

- *Consistency*: the aggregate set of judgments must be consistent, i.e. there is no proposition p such that both p and *not* p are logically derivable from the aggregate set.
- *Completeness*: for every proposition p in the agenda, either p or *not* p is in the aggregate set.
- *Deductive closure*: every proposition p in the agenda which is a logical consequence of the aggregate set is included in the aggregate set.

5.2. Social-choice and judgment-aggregation theoretic analyses of argument aggregation

Let us see now how the current literature deals with these and related properties in the setting of argument aggregation.

Tohmé et al. [71] analyze the aggregation of different abstract attack relations over a common set of arguments. Each of those attack relations can be considered as the representation of a criterion of defeat (or the beliefs of an agent about defeat, etc.). It is well known in the field of SCT that if the conditions of unanimity, anonymity, independence of irrelevant alternatives and non-dictatorship are imposed over an aggregation of preferences, it may be impossible to satisfy them. But a positive result may ensue under the same conditions if the class of winning coalitions (i.e. sets of agents who can impose a choice by voting coordinately) in an aggregation process by voting is a *proper prefilter*. In the case analyzed by these authors, a proper prefilter is such that: 1) the set of all the attack criteria is a winning coalition; 2) every set containing a winning coalition is also a winning coalition; 3) at least one attack criterion belongs to every winning coalition, and 4) no winning coalition has only one attack criterion. The outcome may preserve some features of the competing attack relations, such as the highly desirable property of acyclicity (in terms of Dung, this corresponds to well-founded argumentation frameworks) which can be associated with the existence of a single extension of an argumentation system. The downside of this is that, in fact, the resulting attack relation must be a portion common to all the "hidden dictators" of the system, that is, all the attack criteria that belong to all the winning coalitions (this work is detailed in Section 8.2).

In Rahwan and Larson [65] and its sequels [51,65], agents have *self-interested* preferences, meaning that each agent is only interested in the final status of her own subset of arguments. The status is defined in terms of labellings (see Section 4.2.1). The preferences of the agents are preferences over labellings. An agent has *acceptability-maximizing preferences* if she prefers those labellings that maximize the number of arguments labeled `in` among those of her interest. An interesting property widely studied in SCT is the *Pareto optimality* of outcomes. In this setting, a collective labelling L is Pareto optimal if there is no other labelling L' such that for every agent i, L' is at least as preferred as L and, for some agent j, L' is strictly preferred to L for j. The point is that neither the grounded extension nor the preferred extension(s) correspond in general with labellings that are Pareto optimal. Then the authors investigate the relationship between Pareto optimal outcomes and extension semantics under various sufficient conditions. They prove, for instance, that if agents have acceptability-maximizing preferences then every Pareto optimal labelling corresponds to a preferred extension, and that the grounded extension characterizes exactly the Pareto optimal outcome among a rejection-minimizing population, i.e. consisting of agents that prefer the labellings that minimize the label `out` on their own arguments (more details in Section 8.3).

Other works analyze the possibility of rational collective outcomes under a set of widely accepted social choice principles. Rahwan and Tohmé [69] deal with the strategic aspects of aggregating argument frameworks under *universal domain* (every possible profile is in the domain of the aggregation function), *unanimity*, *anonymity* and *systematicity* (a variant of independence of irrelevant alternatives according to which the collective judgment about two arguments a and b in any two different profiles should be the same if every individual judgment on them is the same in the two profiles). Of particular interest is the ability of manipulating declarations (the same concern as in Rahwan and Larson [65]) in order to obtain a collective framework in which some desired arguments are labeled `in`. An aggregation procedure that disallows this possibility is called *strategy-proof*. It is shown that when voting is restricted to a core set of arguments (the focal set), argument-wise plurality is strategy-proof. On the other hand, as discussed above, this aggregation procedure does not satisfy collective rationality. In this paper this limitation is circumvented when the individual labellings satisfy two very demanding conditions. The first condition, *Condorcet defeat*, indicates that when for any argument a the alternative `out` wins, there must exist another argument b attacking a for which a plurality of agents votes `in`. On the other hand, *non-Condorcet indecision* is obtained if for any argument a the plurality votes yields `in` while none of its attackers can have a plurality of `in` or `undec` votes. Bodanza [11] presents three sensible restrictions on the domain that ensure the collective rationality, at least for the plurality voting mechanism: the number of agents is odd (to avoid ties), the argumentation framework admits only three possible labellings (this leads to the satisfaction of *Condorcet defeat*), and the agents should minimize the number of arguments labeled undec[6] (satisfying thus *non-Condorcet indecision*).

Li [53] establishes an impossibility result for aggregate attack relations, similar to the *Paretian liberal dilemma* in SCT [70]. This dilemma captures the inherent tension between individual rights and collective consensus. The introduction of rights in collective argumentation can be justified, for instance, by the fact that some agents can have a relevant expertise in the subject matter at stake that other agents lack, and thus they are assigned the right to determine the collective defeat relation on some pairs of arguments (cf. Kontarinis, Bonzon, Maudet, and Moraitis [48]). Li shows that an aggregation function yielding a collective defeat relation satisfying the conditions of *universal domain* and *unanimity* is incompatible with the condition of *minimal liberalism*, i.e. the assignment of rights to at least two agents, each of them on at least one pair of arguments.

Some works are concerned with a key problem derived from judgment aggregation, the *Discursive Dilemma* or *Doctrinal Paradox*. This is a widely discussed problem arising in settings where many conflicting judgments are aggregated to yield a collective judgment. More specifically, the paradox occurs when different sets of judgments satisfying desirable rationality conditions are aggregated yielding a collective set of judgments that does not satisfy those same conditions. For instance, consider a consistent set of propositions which are premises from which a particular conclusion can be logically derived. The discursive dilemma arises when a voting mechanism applied both on the premises and the conclusion yields an aggregate set of premises which is inconsistent with the aggregate conclusion. This paradox has motivated important developments in the literature on judgment aggregation (List and Pettit [55]; etc.).

Pigozzi [62] argues that the frequently suggested escape-routes for the discursive dilemma in the judgment aggregation literature, through the so called premise-based and conclusion-based procedures, are not satisfactory methods for group decision-making. She proposes a new aggregation procedure for the case in which the outcome is a set of arguments, combining features of the premise and the conclusion-based procedures. In this setting, an argument is a consistent assignment of truth-values to both the conclusion and its premises, instead of one or another. The premise-based procedure and the conclusion-based procedure are therefore included in this unitary approach that the author calls *argument-based procedure*.

Pigozzi and van der Torre [63] analyze the same problem but looking at the axioms that characterize the aggregation procedure. Noting that majority voting creates also problems in the aggregation of preferences, like the famous *Condorcet Paradox*, in which cycles of preference arise from the aggregation of acyclic individual preferences, they look for conditions that could help to get rid of the paradox. Following List and Pettit [55], they consider four axioms that, given a profile of interpretations $(I_1, \ldots, I_n)$, should be satisfied by the aggregate interpretation I_S, where $S = \{1, \ldots, n\}$: *universal domain*, *collective rationality* (I_S is consistent), *anonymity* and *systematicity* (if in a profile $(I_1, \ldots, I_n)$ a proposition A has the same votes as proposition B in an alternative profile $(I'_1, \ldots, I'_n)$, the truth value or A in I_S should be the truth value of B in I'_S). List and Petit have shown that no aggregation process can satisfy these axioms simultaneously. Pigozzi and van der Torre propose to weaken systematicity, replacing it with a condition they call *premise independence of irrelevant propositional alternatives*, according to which if

[6] This condition means that the labellings are *semi-stable* [19]. It represents here the highest commitment in the decision between in and out.

in a profile $(I_1, \dots, I_n)$ the proposition A has the same votes as in the alternative profile $(I'_1, \dots, I'_n)$, the truth value of $X \in \mathcal{C}(A)$ in I_S should be the same in I'_S, where $\mathcal{C}(A)$ is the set of propositions in the class $\mathcal{L}$ of logical consequences of A. This condition does yield a positive result, namely the existence of consistent ways of aggregating judgments.

Caminada and Pigozzi [20] note that the discursive dilemma can arise in a labelling aggregation setting, since statements such as 'argument a is in', 'argument b is out', etc., are just judgments. They seek ways to ensure the compatibility of a collective outcome with the individual judgments, in the sense that any individual member must be able to defend the collective decision. These authors define a partial ordering $\sqsubseteq$ of labellings: given two labellings $\mathcal{L}$ and $\mathcal{L}'$, $\mathcal{L} \sqsubseteq \mathcal{L}'$ iff $\mathtt{in}(\mathcal{L}) \subseteq \mathtt{in}(\mathcal{L}')$ and $\mathtt{out}(\mathcal{L}) \subseteq \mathtt{out}(\mathcal{L}')$. Several aggregation operators can be defined, avoiding the discursive dilemma. So, for instance, consider the largest admissible labelling $\mathcal{L}_{\mathrm{so}}$ such that $\mathcal{L}_{\mathrm{so}} \sqsubseteq \mathcal{L}_i$ for each labelling in the profile $(\mathcal{L}_1, \dots, \mathcal{L}_n)$. The authors prove that it includes the *skeptical* labelling obtained from the profile (any argument a is labeled in in the aggregate labelling iff a is labeled in in each $\mathcal{L}_i$ and is labeled out iff a is labeled out in each $\mathcal{L}_i$). $\mathcal{L}_{\mathrm{so}}$ is free of the discursive dilemma for argument systems, as well as *credulous* and *super credulous* aggregate labellings obtained in terms of other extension concepts. The credulous aggregate labelling $\mathcal{L}_{\mathrm{co}}$ is motivated by the idea that the group acceptance/rejection of an argument is justified if each individual either agrees on it or, in the worst case, is indifferent. The super credulous aggregate labelling $\mathcal{L}_{\mathrm{sco}}$ is obtained through an "expansion" of the credulous outcome by relabelling illegal undecs to ins and outs. (More details of this work in Section 8.4).

Subsequently, Caminada, Pigozzi, and Podlaszewski [21] analyze the skeptical and credulous aggregation operators from a social welfare perspective (intuitively, through the credulous operator the group assigns in (resp. out) to argument A if there is someone who believes that A is in (resp. out) and nobody thinks that A is out (resp. in), and A is labeled undec in all other cases). The authors study under which conditions these operators are Pareto optimal and whether they are manipulable. To define those conditions it is assumed that individuals have preferences over the possible collective outcomes and the labelling submitted by each agent is her most preferred one (indeed, the one she would like to see adopted by the whole group). Hence, each agent's preference order can be defined according to how "close" is each possible outcome to her own labelling. To this end, the authors define the notions of Hamming set and Hamming distance (see Section 7). The following results are obtained: if individual preferences are either Hamming set or Hamming distance based, then the skeptical 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. Credulous aggregation operators are also Pareto optimal when choosing from the admissible labellings that are compatible with each individual labelling if individual preferences are Hamming set based, but not if they are Hamming distance based.

Booth, Awad, and Rahwan [15] elaborate on this result, pointing out that the aggregate labellings that are obtained by means of *down-admissibility* (the g.l.b. of admissible labellings $L \sqsubseteq L_i$ for each i) as well as of *up-completeness* (the l.u.b. of complete labellings $L_i \sqsubseteq L$) fail to satisfy *independence* (analogous to *Premise Independence of Irrelevant Propositional Alternatives* for labellings and labels in and out). To solve this problem they introduce *interval aggregation*: consider the number of agents that vote out for an argument a, say n_- and those that vote in, n_+. Then (n_-, n_+), belongs to a previously defined family of intervals in S, a is labeled in if $n_+ \geqslant n_-$ and out if $n_+ \leqslant n_-$. If the interval does not belong to the allowed family, a is labeled undec. This procedure satisfies many desirable properties, but it cannot simultaneously satisfy independence and collective rationality. To achieve this, they resort to a labelling $L^{\downarrow}$, which is the g.l.b. among the admissible labellings L such that $L \sqsubseteq L_{\mathtt{int}}$ (where $L_{\mathtt{int}}$ is the aggregate labelling obtained through interval aggregation). Then they obtain the l.u.b. of the complete labellings L such that $L^{\downarrow} \sqsubseteq L$. The resulting labelling $L^{\downarrow\uparrow}$ satisfies collective rationality as well as a weaker form of independence.

Awad et al. [9] note that labelling aggregation, with its tree of "truth values" is closer to 3-value judgment aggregation. In fact, argumentation frameworks are closer to be a representation of contradictory information than of propositional satisfiability. In any case, majority voting is no longer a reasonable procedure when three alternatives are at play. But it can be replaced by plurality voting, in which the alternative with more votes wins. Applied argument-wise, this procedure ensures the satisfaction of several conditions, in particular independence, weak systematicity and non-dictatorship but violates universal domain (it becomes restored when ties are disallowed) and collective rationality. The only way in which this voting procedure yields collective rationality is when the voters are restricted to vote only for arguments in the grounded extension of their respective argument frameworks.

Endriss and Grandi [35] analyze a related problem, namely the aggregation of graphs with more than 3 edges. As a graph is fully determined by its set of edges, a graph aggregator is a function that maps a profile of sets of edges to a set of edges. They show that there does not exist a way to obtain a graph with some desired properties if a graph-theoretic version of independence is required jointly with the non-existence of a dictator, i.e. an individual whose proposed graph becomes always the aggregate one.

Dunne et al. [33] study the computational complexity of checking SCT properties in the context of argument aggregation. They consider Boolean circuits as a "natural" representation of argument aggregation procedures, where the Boolean values $\top$ and $\bot$ stand for the presence and absence of an attack between a pair of arguments, respectively. The complexity of determining whether a given aggregation function has a specific property is reduced to that of finding the value of a propositional formula over some specified basis (e.g. the set of logical connectives $\{\wedge, \vee, \neg\}$). Using this method the authors show, for example, that the problem of checking anonymity is coNP-complete.

Delobelle et al. [27] analyze their distance-based merging operators in terms of the aggregation axioms, as defined in Dunne et al. [33]. These authors prove that these operators satisfy anonymity, identity (if all the argumentation frameworks in the input coincide, the aggregation result must be identical to them), unanimity, and majority (for majoritarian aggregation functions).

Rahwan and Larson [65] present a game-theoretical model of argumentation aggregation in which agents can act strategically revealing or hiding their arguments in such a way that the collective outcome can be manipulated. Their goal is the design of a mechanism which cannot be manipulated by the agents. Then the authors offer a strategy-proof *direct argumentation mechanism* that makes the agents to reveal their arguments simultaneously. Strategy-proofness is guaranteed under the condition that the agents use a skeptical criterion of argument justification (grounded semantics) and their sets of arguments do not include odd-cycles of attack. Rahwan, Larson, and Tohmé [67], moreover, consider more realistic scenarios in which agents can also lie presenting arguments that they do not have in their argument sets. Strategy-proofness is obtained under two conditions: (a) an agent cannot benefit from hiding any of its own arguments, because its arguments cannot "harm" its focal argument, and (b) an agent cannot benefit from revealing any argument it does not have, because it cannot "benefit" its focal argument.

Bonzon, Maudet, and Moretti [14] investigate the uncertainty inherent to debates and apply concepts from cooperative game theory to account for the decision-problem faced by an agent in this context. The authors assume that agents have cardinal preference relations over single arguments, assessing the relevance of each argument with respect to her/his own goals, and facing the uncertainty about the outcome of the debate. The aim is to measure the relative importance (the "worth") of arguments for an agent, taking into account her own preferences as well as the information provided by the attack relations among arguments. The authors develop a game theoretic coalitional model in which a classical power indexes for coalitional games (the Shapley value) is used to measure the impact of a single argument in a debate.

6. Dynamic argument aggregation

Although the dynamic aspects of aggregation are not often made explicit – as we have seen so far – they are certainly implicitly assumed, basically, due to the deliberative character of human argumentation. In fact, the dynamics of argumentation in terms of changes in Dung's argumentation frameworks is studied only in a few works. While previous works included dynamic aspects, they were concerned with the design of specific algorithms in which steps are taken towards a solution (e.g. Bonzon and Maudet [13]). Other works focused on the minimal changes required to fit the collective argumentation framework to some determined goal, for example the expected acceptance or rejection of a subset of arguments. To this end, the literature has found a background in the related field of belief dynamics, mainly in the use of AGM-style postulates [2] that characterize the rationality of revisions. On the other hand, dynamic logics have been addressed using Kripke models, which have also been useful for the characterization of dynamic argumentation [40,41].

Bonzon and Maudet [13] model a debate through a game protocol in which (i) all the agents are focused on the same single issue (argument) of the debate (that is, agents evaluate how good is a state of the debate on the sole basis of the status of this specific argument); (ii) all the agents make use of grounded semantics to evaluate both their private argumentation system and the situation on the common gameboard; and (iii) all the agents share the same set of arguments, but they may have different views on the attack relations between these arguments (e.g., agents can be equipped with value-based argumentation systems ranking differently the values). The agents are divided in two groups, pro and con, and try to impose/refuse the focal argument. In the process, an aggregate argumentation framework is built. At each turn, an agent of a group advances or erases an attack, maybe introducing some new argument (previously introduced arguments cannot be erased). Agents cannot repeat their moves. Once the aggregate argumentation framework is obtained, the surviving arguments are decided using the grounded extension. The authors show that the protocol does not yield, in general, the same argumentation framework as that obtained by merging the individual argumentation frameworks using majority voting on the attacks, but they offer sufficient conditions for the coincidence (this work is detailed in Section 8.5).

Kontarinis, Bonzon, Maudet, Perotti, van der Torre, and Villata [49] study how agents can contribute to a debate in order to reach a goal of accepting or rejecting a specified argument of their interest. The paper focuses on the minimal changes or *target sets* that are required to achieve the goal, in terms of the revision operations of addition or deletion of attacks. The model assumes that in a first phase of the debate the agents reach an agreement about a set of arguments and attacks which are relevant to the subject matter of the debate. The agreement can be reached, for example, through voting. From that point on, the debate focuses on the attacks that have caused disagreement among the agents. It is also assumed that the attacks that are candidates to be accepted or rejected are fixed in the previous phase. In this setting, aggregation questions are relegated to the latter phase. The authors provides some general properties of the set of minimal successful moves and compute target sets by applying rewriting rules written in the Maude programming language.

Kontarinis et al. [48] (a work also commented in Section 7) defines a deliberative process that chooses the right expert on an debated issue. In a first phase, the agents express their opinions, modeled through weighted argumentation systems (*WAS*). In the second stage, an aggregated *WAS* is obtained, and several criteria are used to assess how controversial the outcome is, mainly in terms of the agreement or not on the attacks. If the assessment shows that the *WAS* is too controversial, a third phase is carried out to choose a right expert, based on a dominance order over the available experts. Relevant in this respect is the observation of Polacsek and Cholvy [64], who note that experts can change their minds during a debate. Even accepting this, the latter authors do not include the ensuing dynamic aspects in their proposed framework of debates.

Delobelle et al. [27] work on the expected extensions of the merged aggregation of a profile of argumentation frameworks. The intuition behind this is that merging can be understood in terms of the changes it produces in the argumentation framework itself. They use a language, defined by Coste-Marquis, Konieczny, Mailly, and Marquis [26], to introduce formulas expressing the constraints that a set of arguments must satisfy to become the acceptable outcome of a revision. These possible sets of arguments are intended to be the extensions – for a given semantics – of a final argumentation framework. The revision operators are constructed in a two step-process: the first step is to select the set of revised candidates; the second step is to generate the argumentation frameworks that represent these candidates. The resulting sets of arguments are expected to satisfy some rationality postulates in the style of Katsuno and Mendelzon [45] (in turn, related to AGM). Their process of generation of argumentation frameworks was commented above in Section 4.2.2.

Pedersen and Dyrkolbotn [61] propose a framework to reason about and model deliberation. The authors make use of logical tools, relying both on a truth-functional three-valued view of argumentation [3,34], and on modal logic [40,41]. They focus "on the stepwise, iterative development of a common framework, and on the logical analysis of the different ways in which such a [deliberative] process may unfold, by way of a logical treatment of the modalities that arise from quantifying over the space of possible deliberative futures". The logical language has two levels. The lower level, a Lukasiewicz three-valued logic, is used to refer to static argumentation. The next level, a dynamic modal language, allows to express consequences of updating with a given argument. The semantics of this language is given by a Kripke model, which allows expressing the accessibility relation among states of the world. In this setting, the goal is the determination of the argumentation frameworks that are accessible (through deliberation) from a given argumentation framework describing the actual state of the debate.

7. The collective value of an argument: Using numerical values

To what extent can we say that an opinion differs from another one? In particular, how can such difference be quantified? This question can adopt different specifications in the context of argument aggregation. One of them, for instance, concerns the *distance* between two argumentation frameworks, one representing an individual assessment and the other being the collective outcome. Another – apparently more simple – question is how to quantify the agreement on the acceptance of an argument, a conflict, an attack, etc. This could be specified, for example, just by counting votes. Some papers on argument aggregation propose specifications of such quantitative concepts, employing different techniques.

Coste-Marquis et al. [25,57], as we have discussed in Section 4.1, introduce the idea of using a measure of *distance* between PAFs, defining the merging of a profile as a set of argumentation frameworks that minimizes this distance. The aggregation function is associated to a map from $\mathbb{R}+^n$ (each value $i \leqslant n$ representing the distance of the i-est PAF to the entire profile) to $\mathbb{R}+$, satisfying some properties called non-decreasingness, minimality and identity (more details in Section 8.1).

Rahwan and Podlaszewski [68] seek ways to construct aggregation operators to obtain a set of aggregate labellings up from a profile of individual labellings (see Section 4.2.1). To do that, they combine a technique developed by Miller and Osherson [58] of binary judgment aggregation with a previously defined way of quantifying disagreement among labellings [16], which in turn uses a Hamming distance. Basically, a distance 2 is assigned to "hard" conflicts, i.e. differences in labels `in`/`out`, 1 is assigned to "soft" conflicts, i.e. differences in labels `in`/`undec` and `out`/`undec`, and 0 is assigned if there is no conflict.

Hamming distances are also used by Caminada, Pigozzi, and Podlaszewski [21] to define preference orders among labellings. Given two labellings, the Hamming set between them is the set of all the arguments to which those labellings assign different labels, while their Hamming distance is the cardinality of the Hamming set. These authors prove that some aggregation operators satisfy desirable SCT properties (see Section 5).

Gabbay and Rodrigues [38] also propose a numerical approach to merging argumentation frameworks. They define a process of weight assignation to arguments and attacks. The process can vary according to the specific situation at stake. In particular, the authors develop a dynamics of weights, in which they change until reaching an equilibrium. For example, if a attacks b, and both a and the attack of a on b have a high value, the weight of b decreases.

Pigozzi [62], in a work devoted to solve the *discursive dilemma* (also known as *doctrinal paradox*) in judgment aggregation through argumentation (see Section 5), applies techniques already developed to merge databases, following the lead of Konieczny and Pérez [46]. She considers an additional piece of information given by a class of integrity constraints, that is, rules that have to be respected by both the individual choices and the collective result. Pigozzi postulates a metric between interpretations (closely related to a Hamming metric) and an interpretation satisfying the integrity constraints, minimizing the distance from each of the individual assignments. A consequence is that the chosen interpretation is consistent, i.e. free from the paradox.

In Kontarinis et al. [48] the opinion of the right experts is taken into account to solve conflicts. To do this, the authors define a function that assigns a numerical vector to each attack according to the expertise of the agents that postulate it. The model used is a *weighted argumentation system* (*WAS*) [24,32]. This procedure is a kind of qualified voting mechanism, but the authors do not define a specific protocol for the aggregation process.

In another work, Kontarinis, Bonzon, Maudet, and Moraitis [47] propose a gradual evaluation of arguments in a bipolar argumentation setting. The argumentation is called bipolar since it is based on two relations between arguments. One is the attack relation while the other represents the support given by some arguments to others [22].[7] Kontarinis et al. start considering agents represented by a bipolar argumentation frameworks on the same set of arguments, but with possibly different attack and support relations. Following ideas in [23] the authors propose a *local gradual valuation* of arguments based on three principles: (p1) the valuation of an argument is a function of its direct defeaters and of its direct supporters; (p2) if the quality of the support (resp. defeat) increases then the value of the argument increases (resp. decreases); (p3) if the quality of the supports (resp. defeats) decreases then the quality of the support of an argument decreases (resp. increases). The valuation is modeled through a function on the real interval $[-1, 1]$ and the debate among the agents is on a focal argument. The authors introduce a category-based and a cluster-based protocol, exogenously chosen according to the case at hand. In both cases, the gameboard is a weighted bipolar argumentation framework where the weight is simply the

[7] A bipolar argumentation framework contains a Dung's argumentation framework as a substructure.

difference between the number of agents who favor a given attack and the number of agents who are against it. The collective outcome is obtained by keeping the attacks and support relations with a (strict) majority of agents in favor of them. Agents have preferences with respect to *attraction values* (their interpretation depends on the specific protocol) corresponding to the collective valuations of groups. The debate proceeds by a sequence of moves, consisting on positive assertions of attacks, supports, or rejections of previous assertions. A move is relevant for a group if it makes the valuation closer to the attraction value of the group. The debate ends when the game is stable (no agent can add a relevant move).

Finally, Leite and Martins [52] associate votes to arguments, together with a semantics that assigns each argument a value, drawn from a pre-determined set of possible values, representing the strength of the argument. The authors take into account both the structure of the graph of attacks and the social opinion expressed through votes. Arguments are assumed to have an a priori valuation given by a pair (p, n) of natural numbers, being p and n, respectively, the number of positive and negative votes received by the argument. Then, an aggregation function recursively assigns a real value in the range [0, 1] to the argument, which is determined by both its a priori valuation and the valuations of its attackers. An interesting ingredient of this semantics is that the final values of arguments are not binary (accepted/rejected, in/out, true/false, etc.) but somewhat fuzzy, providing a maybe more realistic representation of how arguments are assessed.

8. Some systems of collective argumentation

In this section we present some prominent systems of collective argumentation that jointly cover the motivation categories deployed in this paper. We have chosen the systems by Coste-Marquis et al. [25,57], Tohmé et al. [71], Rahwan and Larson [65], Caminada and Pigozzi [20], and Bonzon and Maudet [13]. They are presented in chronological order. The first one initiated the study of merging argument systems, presenting also an instance of a framework-wise mechanism and of a distance-based argument assessment method. The second one introduced in the literature the study of social-choice theoretical properties in argument aggregation mechanisms. The third one concerns argumentation in a GT setting, studying strategy-proofness of argument aggregation mechanisms. The fourth one concerns argumentation in a JAT setting, studying the compatibility of a social outcome with the individual judgments in the context of a labelling-based argument-wise aggregation mechanism. Finally, the last one studies dynamical aspects concerning the construction of a common argumentation framework through a debate between two parts.

8.1. Coste-Marquis, Devred, Konieczny, Lagasquie-Schiex, and Marquis: Merging argumentation systems

In [57], the authors offered the first approach to argument aggregation using Dung's argumentation frameworks. Subsequent developments were introduced in [25]. They propose a procedure that starts by making each agent aware of the arguments considered by the rest of the agents. Then each agent expands her framework by incorporating all those arguments. The set of arguments is now common to all the agents, becoming the *status quaestionis* on the basis of which they start discussing. Given that, the agents add new attacks in a consensual way. To do this, each agent keeps her initial attacks and adds all the attacks (a, b) accepted by every agent who had a and b in his original set of arguments. Similarly, if there is a consensus on the non-existence of attacks between a and b, no such attacks are added to the expanded argument systems. In this way, each agent ends up with a *consensual expansion* of her own argumentation framework. Each of these expansions constitutes a *partial argumentation framework* (PAFs), in which, given a pair of arguments a and b, the agent determines either that (a, b) belongs already to her own attack relation or that there is no attack between them or, finally, that she ignores whether an attack exists. More formally, a PAF is a quadruple $\langle A, R, I, N\rangle$ where A is a finite set of arguments and R, I and N are binary relations on A: R is the *attack relation*, I is the *ignorance relation* and is such that $R \cap I = \emptyset$, and $N = A \times A \setminus R \cup I$ is the *non-attack relation* (since N can be deduced from A, R and I, a PAF can be fully specified by a triple $\langle A, R, I\rangle$). A PAF can be "completed" just by transferring any (a, b) from the ignorance class to any of the other two. That is, $AF = \langle A, S\rangle$ is a *completion* of $PAF = \langle A, R, I\rangle$ iff $R \subseteq S \subseteq R \cup I$.

Example 3 ([25])**.** The $PAF = \langle A = \{a, b, c, d\}, R = \{(a, b), (a, c)\}, I = \{(c, a), (b, d)\}\rangle$ has four possible completions:

$$\begin{aligned} AF_1 &= \langle A, R\rangle, \\ AF_2 &= \bigl\langle A, R \cup \bigl\{(c, a)\bigr\}\bigr\rangle, \\ AF_3 &= \bigl\langle A, R \cup \bigl\{(b, d)\bigr\}\bigr\rangle, \\ AF_4 &= \bigl\langle A, R \cup \bigl\{(c, a), (b, d)\bigr\}\bigr\rangle. \end{aligned}$$

Next, a fusion process will select a class of argumentation frameworks representing the profile $\mathcal{P}$. To do this, a pseudo-distance between PAFs is defined, that is, a number which represents how far or close is one PAF to another one. A *pseudo-distance* d between PAFs over A is a mapping which associates a non-negative real number to each pair of PAFs over A and satisfies the properties of symmetry ($d(x, y) = d(y, x)$) and minimality ($d(x, y) = 0$ if and only if $x = y$). The pseudo-distance can be defined in many ways, one of them, for example, being the number of pairs (a, b) that belong to the attack relation of one of the PAFs but do not belong to the attack relation of the other one. Then the authors define the merging of $\mathcal{P}$ as any set of argumentation frameworks that minimizes the distances among the PAFs of the profile. An *aggregation function* is a mapping $\otimes$ from $\mathbb{R}{+}^n$ (each value $i \leqslant n$ representing the distance of the i-est PAF to the entire profile) to $\mathbb{R}+$ that satisfies non-decreasingness (i.e. if $x_i \geqslant x'_i$ then $\otimes(x_1, \dots, x_i, \dots, x_n) \geqslant \otimes(x_1, \dots, x'_i, \dots, x_n)$), minimality (i.e. $\otimes(x_1, \dots, x_n) = 0$ if $\forall i x_i = 0$), and identity (i.e. $\otimes(x) = x$). The merging of a profile is defined as a set of argumentation frameworks. Given a profile $\mathcal{P} = \langle AF_1, \dots, AF_n\rangle$, a pseudo-distance d between PAFs, an aggregation function $\otimes$ and n expansion functions $\mathtt{exp}_1, \dots, \mathtt{exp}_n$, the *merging* of $\mathcal{P}$ is the set of AFs

$$\begin{aligned} &\Delta_d^{\otimes}\bigl(\langle AF_1, \dots, AF_n\rangle, \langle \mathtt{exp}_1, \dots, \mathtt{exp}_n\rangle\bigr) \\ &\quad = \left\{ AF \text{ over } \bigcup_i A_i \mid AF \text{ minimizes } \bigotimes_{i=1}^{n} d\bigl(AF, \mathtt{exp}_i(AF_i, \mathcal{P})\bigr) \right\}. \end{aligned}$$

The contribution of this approach is twofold: on one hand, the authors present a general model for merging argumentation systems; on the other, they define concrete ways in which the model can be instantiated. In the general model, a common set of arguments for all the agents is prescribed; in the instantiation, the union of all the individual sets of arguments is taken, representing the common set obtained in information exchanges. In the general model, an agreement on the attack relation is needed;

in the instantiation, since the union of all the subjective attack relations leads to counterintuitive outcomes, consensual expansions are adopted. In the general model, a merging operator is defined in terms of a distance relation among PAFs; in the instantiation, a pseudo-distance called *edit distance* is introduced, counting the number of pairs of arguments that belong/do not belong to the attack relations of the individual systems. In the general model, an aggregation operator yields the argumentation frameworks closer to the rest of them in a given profile; in the instantiation *sum*, *max* and *leximax* aggregation functions are considered.

8.2. Tohmé, Bodanza, and Simari: Social choice-theoretical analysis of attack aggregation

SCT is concerned with the rationality of aggregation operations on individual preferences. Similarly, the work by Tohmé et al. [71] is concerned with the rationality of aggregation operations on individual attack criteria among arguments. Arrow's Impossibility Theorem [4] claims that four quite natural constraints that capture abstractly the properties of a democratic aggregation process (to wit, the Pareto condition (i.e. unanimity), positive responsiveness, independence of irrelevant alternatives and non-dictatorship) cannot be simultaneously satisfied. That is true for the case of reflexive and transitive preference relations over the alternatives. Once those constraints become incorporated in the framework of argumentation, we could expect something like Arrow's theorem to ensue. But attack relations and preference relations are different in many respects. While preferences are usually formalized as reflexive, transitive and complete relations, attack relations are free to adopt any configuration. This is a reason why an Arrow-like result may not be a necessary outcome for the aggregation of argumentation systems. Indeed, the authors obtain a possibility result, though it is achieved at the cost of accepting the existence of "hidden dictators".

The approach considers an *extended argumentation framework* $AF^n = \langle AR, \rightharpoonup_1, \dots, \rightharpoonup_n \rangle$, for a given n where each $\rightharpoonup_i$ $(1 \leqslant i \leqslant n)$ is a particular attack relation among arguments of AR representing different criteria according to which arguments are evaluated one against another. An *aggregate argumentation framework* is understood as a framework $AF^* = \langle AR, \mathcal{F}(\rightharpoonup_1, \dots, \rightharpoonup_n) \rangle$ where $\mathcal{F}(\rightharpoonup_1, \dots, \rightharpoonup_n) = \rightharpoonup$ is a function of the attack relations of AF^n ($\mathcal{F}$ may be applied over any extended argumentation framework with n attack relations). This is of course analogous to a social system in which a unified criterion must by reached. A paradigmatic instance is that of majority voting, which in this setting can be expressed as follows:

(1) $a \rightharpoonup b$ if $|\{i : a \rightharpoonup_i b\}| > \max(|\{i : b \rightharpoonup_i a\}|, |\{i : b \not\rightharpoonup_i a \wedge a \not\rightharpoonup_i b\}|)$,
(2) $b \rightharpoonup a$ if $|\{i : b \rightharpoonup_i a\}| > \max(|\{i : a \rightharpoonup_i b\}|, |\{i : b \not\rightharpoonup_i a \wedge a \not\rightharpoonup_i b\}|)$,
(3) $(a, b) \notin \rightharpoonup$ (i.e. a does not attack b, nor b does attack a in $\rightharpoonup$) if $|\{i : b \not\rightharpoonup_i a \wedge a \not\rightharpoonup_i b\}| > \max(|\{i : a \rightharpoonup_i b\}|, |\{i : b \rightharpoonup_i a\}|)$.

The authors offer a more or less realistic example:

Example 4 ([71])**.** Consider the following arguments:

a: "Symptoms x, y and z suggest the presence of disease d_1, so we should apply therapy t_1";
b: "Symptoms x, w and z suggest the presence of disease d_2, so we should apply therapy t_2";
c: "Symptoms x and z suggest the presence of disease d_3, so we should apply therapy t_3".

Assume these are the main arguments discussed in a group of three agents (M.D.s), 1, 2 and 3, having to make a decision on which therapy should be applied to some patient. Suppose that each agent i, $i \in \{1, 2, 3\}$, proposes an attack relation $\rightarrow_i$ over the arguments as follows:

$\rightharpoonup_1 = \{(a, b), (b, c)\}$ (agent 1 thinks that it is not convenient to make a joint application of therapies t_1 and t_2 or t_2 and t_3; moreover she thinks that b is more specific than c, hence b defeats c, and that, in the case at stake, symptom y is more clearly present than symptom w. Hence a defeats b),
$\rightharpoonup_2 = \{(a, c), (b, c)\}$ (agent 2 thinks that it is not convenient to apply therapies t_1 together with t_3 or t_2 joint with t_3; moreover she thinks that symptoms y and w are equally present in the case at stake. Furthermore, both a and b are more specific than argument c, hence both a and b defeat c),
$\rightharpoonup_3 = \{(a, c), (c, b)\}$ (agent 3 thinks that it is not convenient to apply t_1 together with t_3 or t_2 with t_3; moreover she thinks that symptom w is not clearly detectable, hence c defeats b, but a is more specific than c, hence a defeats c).

According to majority voting we obtain $\rightharpoonup$ over AR:

$a \rightharpoonup c$, since a attacks c under $\rightharpoonup_2$ and $\rightharpoonup_3$,
$b \rightharpoonup c$, since b attacks c under $\rightharpoonup_1$ and $\rightharpoonup_2$, and
$(a, b) \notin \rightharpoonup$, since $(a, b) \notin \rightharpoonup_2$ and $(a, b) \notin \rightharpoonup_3$.

Another way of aggregating attack relations is by restricting majority voting to a *qualified voting* aggregation function. It fixes a class $U \subset \{\rightharpoonup_1, \dots, \rightharpoonup_n\}$ such that the outcome of majority voting over a pair of arguments is imposed on the aggregate only if every $\rightharpoonup_j \in U$ belongs to the majority (i.e. each member of U has veto power). Otherwise, in the attack relation none of the arguments attacks the other. In the example above, the qualified voting aggregation function with $U = \{\rightharpoonup_2, \rightharpoonup_3\}$ yields $\rightharpoonup = \{(a, c)\}$ (there is no consensus on the interaction between b and c among the members of U). Both majority and qualified voting (with $U \geqslant 2$) satisfy the four mentioned Arrovian properties (Proposition 1, p. 14). Majority voting, on the other hand, is not free of yielding controversial outcomes. One issue is the *Condorcet's Paradox*, i.e. the case in which acyclic individual preference relations lead to cycles in the aggregate relation. Nevertheless, if F is a qualified voting aggregation function and each $\rightharpoonup_i$ is acyclic, then the aggregate $\rightharpoonup$ is acyclic (Proposition 2, pp. 15–16). Though attack relations can adopt any configuration, the focus on acyclic relations is justified by the fact that they characterize well-founded argumentation frameworks, and this has the consequence that all Dung's semantics coincide in a single extension [30].

Every aggregation function (or voting system) can be represented by its class of *decisive sets*, i.e. the winning coalitions that can impose an outcome. In our setting, $U \subseteq \{\rightharpoonup_1, \dots, \rightharpoonup_n\}$ is a *decisive set* iff, if $a \rightharpoonup_j b$ for every $\rightharpoonup_j \in U$ then $a \rightharpoonup b$ (for example, the decisive sets corresponding to majority voting are those coalitions with half of the voters plus one). Then the authors show (Proposition 3, pp. 16–17) that if for every profile $(\rightharpoonup_1, \dots, \rightharpoonup_n)$ of individual acyclic attack relations $\mathcal{F}$ yields an acyclic $\rightharpoonup$, then the Pareto condition, positive responsiveness, independence of irrelevant alternatives and non-dictatorship are guaranteed if and only if the class of decisive sets satisfies the algebraic properties of a *proper prefilter*, to wit: 1) the class formed by all the agents is a decisive set, 2) any set containing a decisive set is also decisive, 3) at least one agent belongs to every decisive set, and 4) no decisive set has exactly one member. More formally, let $\bar{\Omega} = \{\Omega^j\}_{j \in J}$ the class of decisive sets of $\mathcal{F}$, then

1. $\{1, \dots, n\} \in \bar{\Omega}$.
2. If $O \in \bar{\Omega}$ and $O \subseteq O'$ then $O' \in \bar{\Omega}$.
3. Given $\bar{\Omega} = \{\Omega^j\}_{j \in J}$, where $J = |\bar{\Omega}|$, $\bigcap \bar{\Omega} = \bigcap_{j=1}^{J} \Omega^j \neq \emptyset$.
4. No $O \in \bar{\Omega}$ is such that $|O| = 1$.

Though this guarantees that there will be no dictator, some agents will be "hidden dictators" in the sense that while these agents cannot impose a result, they still have veto power. So, the collectively chosen attacks will be part of the attacks chosen by those agents (Proposition 4, pp. 18–19).

Finally, if each $\rightharpoonup_i$ is acyclic and $\mathcal{F}$ is such that its class of decisive sets form a proper prefilter, then the resulting collective attack relation $\rightharpoonup$ is acyclic, determining a unique extension which is grounded, preferred and stable (Proposition 5, p. 19).

8.3. Rahwan and Larson: Game-theoretic argumentation mechanism design

In [65], Rahwan and Larson observe that the outcome of argumentation is determined not only by the rules of argument acceptability, but also by the strategies employed by the agents who present these arguments. Since agents can be self-interested and may have different preferences over the arguments, the design of the argument acceptability rule should take the mechanism design perspective (a well-known concept from Economics and GT). The authors pose the question: what game rules guarantee a desirable social outcome when each self-interested agent selects the best strategy for itself? In a multi-agent setting where agents introduce the arguments, manipulation can arise if by hiding an argument an agent can change the status of another argument, such that the result is better according to her own interests. Then Rahwan and Larson apply the tools of GT and mechanism design to abstract argumentation frameworks, finding argument evaluation criteria ensuring strategy-proofness.

A set of self-interested agents is denoted by I. The *type* $\theta_i \in \Theta$ of agent i is drawn from the set of possible types Θ. The type represents the preferences and private information of an agent. The preferences of an agent are defined over a set $\mathcal{O}$ of possible *outcomes*. In the context of an argumentation framework $AF = \langle AR, \rightharpoonup \rangle$, the type of agent i, $A_i \in AR$, is the set of arguments that the agent is capable of putting forward (note that agents can operate on the set of arguments but not on the attack relation, which is fixed). Given the agents' types (argument sets) a social choice function f maps a type profile into a subset of arguments, $f : 2^{AR} \times \cdots \times 2^{AR} \to 2^{AR}$. Given a semantics $\mathcal{S}$, and given a type profile $(A_1, \dots, A_I)$, the *argument acceptability social choice function* f is defined as the set of acceptable arguments given the semantics $\mathcal{S}$, in symbols, $f(A_1, \dots, A_I) = Acc(\langle A_1 \cup \cdots \cup A_I, \rightharpoonup \rangle, \mathcal{S})$. The set of possible outcomes is $\mathcal{O} = 2^{AR}$, and the preferences of the agents are expressed using utility functions u where $u_i(o, A_i)$ denotes agent i's utility of outcome o when her type is the argument set A_i.

Agents may not have incentive to reveal their true type because they are able to influence the final argument status assignment by lying, and thus obtain higher utility. There are two ways that an agent can lie in the model. On one hand, an agent might create new arguments that it does not have in her argument set. Another way is by hiding some of her arguments. Either by presenting arguments that the agent does not subscribe or by refusing to reveal certain arguments, an agent might be able to break defeat chains in the argumentation framework, thus changing the final set of acceptable arguments. A strategy of an agent specifies a complete plan that describes what action the agent takes at every stage at which she has to make a decision. In this model, the actions available to an agent involve announcing sets of arguments. Thus a *strategy* $s_i \in \Sigma_i$ for agent i would specify, for each possible subset of arguments that could define its type, what set of arguments to reveal. Then, a *direct argumentation mechanism* is defined as $\mathcal{M}^{\mathcal{S}}_{AF} = (\Sigma_1, \dots, \Sigma_I, g(\cdot))$, where $\Sigma_i = 2^{A_i}$ and $g : \Sigma_1 \times \cdots \times \Sigma_I \to 2^{AR}$. That is, the mechanism determines a subset of acceptable arguments for each profile of agents' strategies (revealed arguments).

Consider now that agents have *acceptability maximizing preferences* in the sense that, given two possible outcomes o_1 and o_2, if $|A_i \cap o_1| \geqslant |A_i \cap o_2|$ then $u_i(o_1, A_i) \geqslant u_i(o_2, A_i)$ for every agent i (i.e.

agents prefer those outcomes that maximize the number of arguments of their respective types). And consider a *skeptical* direct argumentation mechanism, to wit, a mechanism in which $\mathcal{S}$ is the grounded semantics. The following example shows how the mechanism can be manipulated.

Example 5 ([65]). Consider a skeptical direct argumentation mechanism with three agents x, y and z with types $A_x = \{a_1, a_4, a_5\}$, $A_y = \{a_2\}$ and $A_z = \{a_3\}$, respectively. And suppose that the attack relation is defined as follows: $\rightharpoonup = \{(a_1, a_2), (a_2, a_3), (a_3, a_4), (a_3, a_5)\}$. If each agent reveals its true type, then we obtain the argumentation framework $\langle\{(a_1, a_2, a_3, a_4, a_5)\}, \{(a_1, a_2), (a_2, a_3), (a_3, a_4), (a_3, a_5)\}\rangle$, and then the mechanism outcome rule produces the outcome $o = \{a_1, a_3\}$. If agents have individual acceptability maximizing preferences, with utilities equal to the number of arguments accepted, then $u_x(o, \{a_1, a_4, a_5\}) = 1$; $u_y(o, \{a_3\}) = 1$; and $u_z(o, \{a_2\}) = 0$. But then, if x reveals $\{a_4, a_5\}$ instead of $\{a_1, a_4, a_5\}$, then the resulting argumentation framework is $\langle\{(a_2, a_3, a_4, a_5)\}, \{(a_2, a_3), (a_3, a_4), (a_3, a_5)\}\rangle$ with outcome $o' = \{a_2, a_4, a_5\}$, which yields the utility $u_x = 2$.

The authors find a sufficient condition to make the skeptical direct argumentation mechanism strategy-proof when agents have acceptability maximizing preferences: each agent's type is a conflict-free set of arguments which does not include indirect attacks (i.e. odd-length sequences of attack among its members) (Theorem 32, p. 1037). Conversely, strategy-proof mechanisms implementing skeptical social choice functions lead to agents' strategies free of indirect attacks (Theorem 33, ibid.).

Subsequently, in [66] and [51], Rahwan and Larson focus on the property of Pareto optimality, which measures whether an outcome can be improved for one agent without harming other agents. In this setting, the authors change the methodology from using extension semantics to using labellings. These works were commented in Section 5.2.

8.4. *Caminada and Pigozzi: Judgment aggregation in abstract argumentation*

In Caminada and Pigozzi [20], judgment aggregation is studied as the aggregation of individual labellings of a given argumentation framework. It constitutes the first approach to judgment aggregation from the point of view of argumentation theory, and the first in using abstract argumentation to that end. The authors are not particularly concerned with the discursive dilemma – a central subject of JAT – but with defining social outcomes that any individual participating in the decision could subscribe while guaranteeing collective rationality. Given a set of individuals $N = \{1, \dots, n\}$, the aim is to define a general labelling aggregation operator F_{AF} that assigns a collective labelling $\mathcal{L}_{Coll}$ to each profile $\{\mathcal{L}_1, \dots, \mathcal{L}_n\}$ of individual labellings. Note that the profiles here are defined as sets instead of as vectors. This is because though several individuals may submit the same judgments set (i.e. the same labellings), the motivation is to avoid situations in which any group member is forced to commit herself to a position that goes against her opinion. Therefore, cardinality considerations (regarding, for instance, minoritarian vs. majoritarian opinions) do not play any role in the authors' approach. Consequently, a *general labellings aggregation operator* is defined as a function F_{AF} such that, given an argumentation framework $AF = \langle AR, \rightharpoonup\rangle$ and the set $\mathcal{L}abellings$ of all the possible labellings of AF, $F_{AF} : 2^{\mathcal{L}abellings} \setminus \{\emptyset\} \rightarrow \mathcal{L}abellings$ and $F_{AF}(\{\mathcal{L}_1, \dots, \mathcal{L}_n\}) = \mathcal{L}_{Coll}$. Then the authors define two kinds of outcomes: skeptical outcomes and credulous outcomes.

The skeptical outcome is illustrated as follows. Assume all the group members are gathered in a meeting with a chair who asks the opinion of the participants about each argument. Then the authors explain: "If all participants unanimously think the argument should be accepted, then the argument is initially accepted. If all participants unanimously think the argument should be rejected, then the argument is initially rejected. In all other cases, the group as a whole does not have an explicit opinion about the argument. After all arguments have been treated this way, the meeting goes to the second phase. The chairman evaluates whether each acceptation or rejection can be justified by the outcome found so far. An argument that is accepted without every defeater being rejected can no longer be accepted. Conversely, an argument rejected without an accepted defeater can no longer be rejected. In each of these two cases, the group has to abstain from stating any further opinion about those arguments. This is an iterative process, since once one abstains from stating an explicit opinion about a particular argument, opinions (acceptation or rejection) about related arguments can be no longer justified. The process does not stop until the group no longer has unjustified explicit opinions. After this second phase is over, the result will be a position that is "smaller or equal" (less or equally committed) than each individual position of the participants. That is, each argument that is accepted by the group is also accepted by each individual participant, and each argument that is rejected by the group is also rejected by each individual participant." (p. 77). In this way, the skeptical outcome is the most committed position. Let us see the authors' formalisms to capture these ideas. Given two labellings $\mathcal{L}_1$ and $\mathcal{L}_2$, $\mathcal{L}_1$ is *less or equally committed as* $\mathcal{L}_2$ ($\mathcal{L}_1 \sqsubseteq \mathcal{L}_2$) iff $\mathtt{in}(\mathcal{L}_1) \subseteq \mathtt{in}(\mathcal{L}_2)$ and $\mathtt{out}(\mathcal{L}_1) \subseteq \mathtt{out}(\mathcal{L}_2)$. The relation $\sqsubseteq$ is a partial order on labellings (i.e. reflexive, transitive and antisymmetric). As a result, the set of admissible labellings that are smaller or equal to a given labelling $\mathcal{L}$ has a (unique) biggest element (Theorem 5, p. 78). The biggest element of the set of all the admissible labellings that are less or equally committed as a labelling $\mathcal{L}$ is called the *down-admissible* labelling of $\mathcal{L}$. Then, once the group has found the skeptical initial labelling $\mathcal{L}_{\mathrm{sio}}$ in the first phase of the meeting, the second phase consists in finding the down-admissible labelling of $\mathcal{L}_{\mathrm{sio}}$. The *skeptical initial labelling* is found through the operator $\mathrm{sio}_{AF} : 2^{\mathcal{L}abellings} \setminus \{\emptyset\} \to \mathcal{L}abellings$ such that $\mathrm{sio}_{AF}(\{\mathcal{L}_1, \ldots, \mathcal{L}_n\}) = \{(a, \mathtt{in}) \mid \forall i \in \{1, \ldots, n\} : \mathcal{L}_i(a) = \mathtt{in}\} \cup \{(a, \mathtt{out}) \mid \forall i \in \{1, \ldots, n\} : \mathcal{L}_i(a) = \mathtt{out}\} \cup \{(a, \mathtt{undec}) \mid \exists i \in \{1, \ldots, n\} : \mathcal{L}_i(a) \neq \mathtt{in} \wedge \mathcal{L}_i(a) \neq \mathtt{out}\}$.

To find the down-admissible labelling of $\mathcal{L}_{\mathrm{sio}}$ the authors define a series of contractions. First, a contraction function $c_{AF} : \mathcal{L}abellings \times AR \to \mathcal{L}abellings$ is such that $c_{AF}(\mathcal{L}, a) = (\mathcal{L} \setminus \{(a, \mathtt{in}), (a, \mathtt{out})\}) \cup \{(a, \mathtt{undec})\}$. Second, given a labelling $\mathcal{L}$, a *contraction sequence from* $\mathcal{L}$ is a list of labellings $[\mathcal{L}_1, \ldots, \mathcal{L}_m]$ such that:

(1) $\mathcal{L}_1 = \mathcal{L}$,
(2) for each $j \in \{1, \ldots, m\}$: $\mathcal{L}_{j+1} = c_{AF}(\mathcal{L}_j, a)$, where a is an argument that is illegally in or illegally out in $\mathcal{L}_j$, and
(3) $\mathcal{L}_m$ is a labelling without any illegal in or illegal out.[8]

Then, the authors show that the down-admissible labelling $\mathcal{L}_{da}$ of $\mathcal{L}$ coincides with the last element $\mathcal{L}_m$ of a contraction sequence from $\mathcal{L}$ (Theorem 6, p. 80). The contraction sequence serves to calculate a *skeptical aggregation operator*, characterized as $\mathrm{so}_{AF} : 2^{\mathcal{L}abellings} \setminus \{\emptyset\} \to \mathcal{L}abellings$ such that $\mathrm{so}_{AF}(\{\mathcal{L}_1, \ldots, \mathcal{L}_n\})$ is the down-admissible labelling of $\mathrm{sio}_{AF}(\{\mathcal{L}_1, \ldots, \mathcal{L}_n\})$. Then $\mathcal{L}_{\mathrm{so}} = \mathrm{so}_{AF}(\{\mathcal{L}_1, \ldots, \mathcal{L}_n\})$ is the biggest admissible labelling such that for every $i \in N$, $\mathcal{L}_{\mathrm{so}} \sqsubseteq \mathcal{L}_i$ (Theorem 7, p. 82. This result also holds for the case of complete labellings: Theorem 8, p. 83). The authors thus capture their intuition: "One positive feature of the thus described sceptical aggregation is that if an argument is accepted (or rejected) by the group outcome, it is also accepted (or rejected) by each individual member of the group. Hence, if a member needs to explain (perhaps in public) why his group

[8] A labelling $\mathcal{L}(a)$ is illegal if if it does not match the conditions stated for complete labellings.

accepts or rejects a particular argument, he will be able to do so without having to go against his own private opinions." (p. 82). On the other hand, the above results have a cost: either unanimity, i.e. the property by which a labelling subscribed by each agent must be the collective one, or the admissibility of the collective labelling have to be resigned.

As a counterexample the authors present the case of a *floating defeat*:

Example 6 ([20])**.** Let $AF = \langle\{a, b, c, d\}, \{(a, b), (b, a), (a, c), (b, c), (c, d)\}\rangle$, $N = \{1, 2\}$. Both $\mathcal{L}_1 = (\{a, d\}, \{b, c\}, \{\})$ and $\mathcal{L}_2 = (\{b, d\}, \{a, c\}, \{\})$ are admissible labellings, but $\mathrm{sio}_{AF}(\mathcal{L}_1, \mathcal{L}_2) = (\{d\}, \{c\}, \{a, b\})$ is not admissible.

The credulous outcome is motivated by the idea that the group acceptance/rejection of an argument is justified if each individual either agrees on it or, in the worst case, is indifferent. A *compatibility* relation among labellings is defined: $\mathcal{L}_1 \approx \mathcal{L}_2$ iff $\mathtt{in}(\mathcal{L}_1) \cap \mathtt{out}(\mathcal{L}_2) = \emptyset$ and $\mathtt{out}(\mathcal{L}_1) \cap \mathtt{in}(\mathcal{L}_2) = \emptyset$ (this relation is reflexive, symmetric and non-transitive. Moreover, if $\mathcal{L}_1 \sqsubseteq \mathcal{L}_2$ then $\mathcal{L}_1 \approx \mathcal{L}_2$). The process of finding the aggregate labelling resembles the skeptical one, but in this case an opinion of the kind in/out finds collective acceptance if each individual either agrees on it or abstains subscribing undec. The *credulous initial labelling* is found through the operator $\mathrm{cio}_{AF} : 2^{\mathcal{L}abellings} \setminus \{\emptyset\} \rightarrow \mathcal{L}abellings$ such that $\mathrm{cio}_{AF}(\{\mathcal{L}_1, \ldots, \mathcal{L}_n\}) = \{(a, \mathtt{in}) \mid \exists i \in \{1, \ldots, n\} : \mathcal{L}_i(a) = \mathtt{in} \wedge \nexists i \in \{1, \ldots, n\} : \mathcal{L}_i(a) = \mathtt{out}\} \cup \{(a, \mathtt{out}) \mid \exists i \in \{1, \ldots, n\} : \mathcal{L}_i(a) = \mathtt{out} \wedge \nexists i \in \{1, \ldots, n\} : \mathcal{L}_i(a) = \mathtt{in}\} \cup \{(a, \mathtt{undec}) \mid \forall i \in \{1, \ldots, n\} : \mathcal{L}_i(a) = \mathtt{undec} \vee (\exists i \in \{1, \ldots, n\} : \mathcal{L}_i(a) = \mathtt{in} \wedge \exists i \in \{1, \ldots, n\} : \mathcal{L}_i(a) = \mathtt{out})\}$. The credulous initial labelling $\mathcal{L}_{\mathrm{cio}} = \mathrm{cio}_{AF}(\{\mathcal{L}_1, \ldots, \mathcal{L}_n\})$ is such that for every $i \in N$, $\mathcal{L}_{\mathrm{cio}} \approx \mathcal{L}_i$ (Theorem 9, p. 85). The *credulous aggregation operator* co_{AF} is defined only on the domain of admissible labellings and it also only returns admissible labellings. As a result, the ensuing credulous aggregate labelling $\mathcal{L}_{\mathrm{co}}$ is such that for every $i \in N$, $\mathcal{L}_{\mathrm{co}} \approx \mathcal{L}_i$ (Theorem 10, p. 86. In this case, the result cannot be extended to complete labellings).

The authors then apply a similar methodology to define a *super-credulous aggregation operator*. We make an informal presentation here. The idea is to apply an *expansion function* to "expand" the credulous outcome by relabelling illegal undecs to ins and outs. A third phase in the meeting is conceived in which each undecided argument is accepted if all its defeaters were previously rejected, and rejected if some of its defeaters was previously accepted. The process goes on (an *expansion sequence* is defined) until every argument that can be accepted is accepted and each argument that can be rejected is rejected (no illegal undec remains). It is proved that the ensuing sequence meets the super-credulous outcome $\mathcal{L}_{\mathrm{sco}}$ (Theorem 12, p. 89), and that for every $i \in N$, $\mathcal{L}_{\mathrm{sco}} \approx \mathcal{L}_i$ (Theorem 13, p. 90).

In sum, this work proposes particular aggregation mechanisms oriented to get collective outcomes that any individual can defend without betraying her own opinions. Moreover, the collective labellings are guaranteed to keep rationality in terms of admissibility and related argumentation-theoretical semantical concepts.

8.5. Bonzon and Maudet: On the outcomes of multiparty persuasion

In [13], Bonzon and Maudet study dynamical aspects of argumentation and propose a multiparty persuasion protocol. Close in spirit to the work of Rahwan and Larson [65,66], the authors consider that agents have a chance to influence the outcome of an argumentation game depending on how they play, and use game-theoretical concepts for their study. They also take for granted that agents' argument moves should immediately improve their satisfaction with respect to the current situation of the debate.

In this context, a number ($n > 2$) of non-coordinated agents exchange arguments on a common gameboard (among the motivating applications they have in mind are, for example, online platforms allowing users to asynchronously modify the content of a collective debate). To keep things simple, the following assumptions are made: (i) all the agents are focused on the same single argument; (ii) all the agents make use of the same argumentation semantics, to wit, grounded semantics, to evaluate both their private argumentation system and the situation on the common gameboard; and (iii) all the agents share the same set of arguments, but they may have different views on the attack relations between these arguments (giving raise to a framework-wise approach).

The model considers a set N of agents, each one holding an argumentation framework $AF_i = \langle AR, \rightharpoonup_i \rangle$, sharing the same arguments AR. To tackle the problem of the collective view, a notion of merged system is considered, adopting a majority voting approach. Ties are broken in favor of the absence of an attack, which allows to ensure the existence of a single merged argumentation system $MAS_N = \langle AR, \rightharpoonup \rangle$, where $(a, b) \in \rightharpoonup$ iff $|\{i \in N : (a, b) \in \rightharpoonup_i\}| > |\{i \in N : (a, b) \notin \rightharpoonup_i\}|$. The corresponding merged outcome is the grounded extension ($\mathcal{E}$) of MAS_N, denoted $\mathcal{E}(MAS_N)$. Agents contribute to the debate in a step-by-step process, guided by their individual assessment of the current state of the discussion, and without coordination with the other agents. The debate has a focused argument, called the *issue d* of the debate, and agents are concerned with the coincidence of the acceptability status of d in the merged system and in their individual systems. Then the debate can be seen as opposing two groups of agents, $PRO = \{i : d \in \mathcal{E}(AF_i)\}$ and $CON = \{i : d \notin \mathcal{E}(AF_i)\}$. The gameboard is a "common" *weighted* argumentation system, where the weight is simply a number equal in the difference between the number of agents who asserted a given attack and the number of agents who opposed it. The fact that the attack of argument a on argument b has a weight α is denoted $a \rightharpoonup_\alpha b$. Let $A(GB)$ be the set of all the arguments present on the gameboard. The collective outcome is obtained by applying the semantics used on the argumentation framework $\langle A(GB), \rightharpoonup \rangle$, where $\rightharpoonup \in A(GB) \times A(GB)$ and $\rightharpoonup = \{a \rightharpoonup_\alpha b : \alpha > 0\}$. In words, the attacks retained are only those supported by a (strict) majority of agents having expressed their view on this relation.

The protocol proceeds in rounds which alternate between the two groups PRO and CON. Within these groups though, no coordination takes place: the agents may for instance play asynchronously and the authority simply picks the first permitted and relevant move before returning the token to the other side. Permitted moves are simply positive assertions of attacks $a \rightharpoonup b$ (with b being an element of the set of arguments $A(GB)$ at round t, $A^t(GB)$), or contradiction of (already introduced) attacks (with (a, b) being an element of the attack relation $\rightharpoonup$ at round t, $\rightharpoonup^t$). A move is *relevant* at round t for a PRO agent (resp. CON agent) if it puts the issue back in (resp. drops the issue from) $\mathcal{E}(AF^t(GB))$. Furthermore, the protocol prevents the repetition of similar moves from the same agent.

Then the protocol is defined as follows:

(1) Agents report their individual view on the issue to the central authority, which then assigns (privately) each agent to PRO or CON.
(2) The first round starts with the issue on the gameboard and the turn given to CON.
(3) Until a group of agents cannot move, we have:
   - agents independently propose moves to the central authority;
   - the central authority picks the first (or at random) relevant move from the group of agents whose turn is active, update the gameboard, and passes the turn to the other group.

When a (relevant) move is played on the gameboard, the following update operation takes place:

(1) after an assertion $a \rightharpoonup b$

- if $a \rightharpoonup_\alpha b \in \rightharpoonup^t$ then $\alpha := \alpha + 1$;
- if $a \rightharpoonup_\alpha b \notin \rightharpoonup^t$ and $a, b \in A^t(GB)$ then the edge is created with $\alpha := 1$;
- otherwise (a is not present), then the node of the new argument is created and the edge is created with $\alpha := 1$;

(2) after a contradiction of $a \rightharpoonup b$, we have $\alpha := \alpha - 1$.

When (after a sequence σ of moves) a group of agents cannot move, the gameboard is said to be stable, and the outcome of the merged system $\mathcal{E}(AF(GB))$ is reached. The protocol satisfies the properties of termination (provided the argumentation frameworks are finite), guaranteed convergence to the merged outcome, and reachability of the merged outcome.

Example 7. Let 1, 2, and 3 be three agents with their argumentation frameworks AF_1, AF_2 and AF_3, respectively, and the merged argumentation framework MAS as follows:

$$\begin{aligned}
AF_1 &= \big\langle \{a, b, c\}, \big\{(a, b), (a, c)\big\}\big\rangle; & \mathcal{E}(AF_1) &= \{a\}\\
AF_2 &= \big\langle \{a, b, c\}, \big\{(a, b), (b, c)\big\}\big\rangle; & \mathcal{E}(AF_2) &= \{a, c\}\\
AF_3 &= \big\langle \{a, b, c\}, \big\{(b, c)\big\}\big\rangle; & \mathcal{E}(AF_3) &= \{a, b\}\\
MAS &= \big\langle \{a, b, c\}, \big\{(a, b), (b, c)\big\}\big\rangle; & \mathcal{E}(MAS) &= \{a, c\}
\end{aligned}$$

The issue of the dialogue is the argument c. We have $CON = \{1, 3\}$ and $PRO = \{2\}$. At the beginning, we have $AF^0(GB) = \langle\{c\}, \{\}\rangle$ with $\mathcal{E}(AF^0(GB)) = \{c\}$. A sequence of moves allowed by the protocol is the following (the second column records the provisional attack relation at each time t):

$t = 1$: agent 1 plays for CON introducing $a \rightharpoonup c$;	$a \rightharpoonup c$
$t = 2$: agent 2 plays for PRO removing $a \rightharpoonup c$;	–
$t = 3$: agent 3 plays for CON introducing $b \rightharpoonup c$;	$b \rightharpoonup c$
$t = 4$: agent 2 plays for PRO introducing $a \rightharpoonup b$;	$a \rightharpoonup b, b \rightharpoonup c$
$t = 5$: agent 3 plays for CON removing $a \rightharpoonup b$;	$b \rightharpoonup c$
$t = 6$: agent 2 cannot move;	$b \rightharpoonup c$

At this point the gameboard is stable and the outcome $\mathcal{E}(AF(GB)) = \{a, b\}$ is obtained.

Note that the status of an issue in the merged argumentation system can contradict the opinion of the majority. Bonzon and Maudet agree with Coste-Marquis et al. that if agents vote on extensions, a lot of significant information is not exploited. (Moreover, this example replicates the commutation problem discussed in Section 4.3.)

In order to characterize the status of the issue obtained by the protocol the authors introduce the notion of global arguments-control graph (ACG). The idea is to gather the attacks of all agents in the same argumentation graph, and then determine which group, PRO or CON, has the control over some path of this graph, and thus a possible way to reach its preferred outcome. Let $add_{(a,b)} = \{i \in N : (a, b) \in \rightharpoonup_i\}$ and $rem_{(a,b)} = \{i \in N : (a, b) \notin \rightharpoonup_i\}$. Then $X \in \{PRO, CON\}$ has *constructive control* of (a, b) in $\rightharpoonup = \bigcup_{i=1}^{n} \rightharpoonup_i$ iff the number of agents in X who can add (a, b) is greater than the number of agents in $\bar{X} \in \{PRO, CON\} \setminus \{X\}$ who can remove it. $X \in \{PRO, CON\}$ has *destructive control* of (a, b) in

$\rightharpoonup = \bigcup_{i=1}^{n} \rightharpoonup_i$ iff the number of agents in X who can remove (a, b) is greater than the number of agents in $\bar{X} \in \{PRO, CON\} \setminus \{X\}$ who can add it. Observe that the notion of destructive control intuitively says that a group has the control to overturn any possible attempt to establish a given relation. Then the *global arguments-control graph* $ACG_N = \langle AR, \rightharpoonup \rangle$ is constructed as follow: (1) $\rightharpoonup = \bigcup_{i=1}^{n} \rightharpoonup_i$, (2) label each $(a, b) \in \rightharpoonup$ by the information about control and playability for each group $X \in \{PRO, CON\}$ ((a, b) is *playable* by X if $(a, b) \in \rightharpoonup_i$ for some $i \in X$).

The authors establish some properties and then show conditions under which the *MAS* is reachable (though convergence is not guaranteed).

9. Other works

In this section we want to mention some related works that we have not discussed before, because they are either not strictly about a collective outcome in argument aggregation or because they are not based on Dung's argumentation frameworks.

Ontañón and Plaza [59] aim to model a deliberation process among learning agents that decide upon some joint decision. They define a specific protocol for the aggregation of attack criteria. Agents have different data about a specific case, but they use a similar criterion of argument evaluation calculating favorable cases over total cases. The aggregation proceeds according to a protocol specifying that each agent proposes a solution to a query. The agents make, in turns, local evaluations of the solutions. Each agent looks for counterarguments (or counterexamples) to the solutions proposed by the other agents and make them public. The new information is incorporated by the agents who reevaluate the arguments, leading them to accept or reject some of them (including some of her own). The process goes on until either a unanimous solution is found or no new counterarguments are presented, in which case a voting mechanism is applied. The agents use case-based reasoning to learn from past cases (where a case is a situation and its outcome) in order to predict the outcome of a new situation. In another paper [60], the authors use this technique in order to generate predictions in a *predictions market*. Each agent has a base case, which consists of pairs of cases and solutions. Agents try to obtain information from other agents in order to increase the accuracy of their predictions. Arguments are generated by predicting a solution to a given case, based on other known cases and solutions. Similarly, counterarguments and counterexamples can be generated. Following the protocol, agents can either assert an argument supporting a prediction or reject a prediction with a counterargument/counterexample.

Grosse, Chesñevar, Maguitman, and Estevez [39] contend that the information made available in Twitter can be useful for modeling arguments which emerge bottom-up from the social interaction associated with such messages, thus enabling an integration between Twitter and defeasible argumentation. The authors outline the main elements characterizing this integration in the context of a particular e-government platform (Decide 2.0). They propose a method for building arguments from aggregated opinions, leading to the design of a platform that makes it possible to explore collective opinions in a more meaningful and systematic manner.

Dignum and Vreeswijk [28] propose a test bed for multi-party dialogues. They combine computational dialectics with a blackboard system to build a "dialectic blackboard architecture". Because multi-party dialogues require different methods of communication than in the case of one-to-one and one-to-many communication modes, the authors use the blackboard system metaphor as a basis for communication. A *discussion group* is formalized as a tuple $\langle G, F \rangle$, where G is a possibly infinite set of agents that communicate by means of a forum F, which consists of an array of messages. Though messages are not thought particularly as arguments, the authors propose to include, as a further step, persuasion dialogues, giving the agents the capability to deal with arguments and counterarguments. Several issues about multi-party dialogues are explored: open vs. closed systems, roles of the parties, medium and addressing, coordination, termination as well as internal operation.

Finally, experimental assessments of argument aggregation mechanisms have began to be explored. Awad, Bonnefon, Caminada, Malone, and Rahwan [8] show that formal models do not capture a number of factors that people consider important in real-life applications. In their work, the authors contrast two aggregation mechanisms, argument-wise plurality rule (AWPR) and skeptical and super credulous operators (SSCOs), the first one based on the idea that an assessment of an argument (in/out/undecided) is chosen if it is submitted by the majority regardless what the minority think, and the later one supporting the idea that minority's opinion should not be completely ignored. The results suggest that AWPR is the generally preferred rule, except in situations where the decision may inflict personal harm to an individual, and the resolution is passed by a narrow margin. The presence of any of these two factors decreases the preference for AWPR, and their joint presence led people to hesitate between AWPR and SSCOs. On the other hand, the authors contrasted the hypothesis that SSCOs would be preferred in situations where all committee members have to defend or take responsibility for the committee's collective decision. The observed result is that preference for SSCOs does not increase in these situations.

10. Conclusions

Different models surveyed in this paper deal with different aspects of the collective argumentation problem, so there is no clear-cut way of comparing them. Moreover, different motivations and techniques are usually combined in the same model. In Table 2 we offer a rough categorization of the works in the literature according to motivational issues and approaches (note that some papers appear simultaneously in different categories).

The debate about which kind of mechanism, the argument-wise or the framework-wise, is more reasonable deserves a deeper philosophical analysis that has yet to be carried out. While such analysis can be done in terms of general foundational principles, we believe that such debate can be carried out in more pragmatic terms, so the reasonableness of one approach or the other will depend on the specific collective argumentation context under consideration (e.g. the framework-wise approach could be the

Table 2

Guidance table of the works on collective argumentation

Motivational issue	Approach		Works
Aggregation mechanisms	framework-wise		[1,6,7,12,18,25,27,33,35,37,38,57,61,71,72]
	argument-wise	labellings	[9,11,17,20,66,69]
		extensions	[6,7,12,27]
Rationality properties of aggregation	SCT, JAT, GT		[9,13,15,20,21,27,33,35,51,53,61–63,65,66,69,71]
Dynamic argumentation	protocols, algorithms		[13,47,48]
	logics, rationality postulates		[27,61]
Social argument assessment	distances		[16,21,25,57,62,68]
	weights, strengths, etc.		[38,47,48,52]
Other	miscellaneous		[8,28,39,59,60,64]

most reasonable one in the context of deliberative democracy, while the argument-wise approach could be the most reasonable one in the context of a debate among experts).

The interaction of collective argumentation with SCT, JAT and GT stands out, in our opinion, as the most promising approach in the field. While research along these lines has been developed for years, there are a lot of techniques and results that can still be profitably exploited in the context of argument aggregation. In particular, we hope for a wider and deeper relationship between collective argumentation and game theoretical aspects related to mechanism design. This, in turn, will help to connect collective argumentation with the relatively new area of computational social choice, and with economic theory in general.

On the other hand, the incipient works on dynamical aspects of collective argumentation, in particular, the design of protocols and algorithms, show the potential contributions of this research line in concrete and practical applications to diverse areas like multi-agent systems, e-government, e-democracy, opinion-mining, prediction markets, social marketing intelligence, etc.

Acknowledgement

This work was supported by Agencia Nacional de Promoción Científica y Tecnológica, Argentina [PICT 2013-1489].

References

[1] S. Airiau, E. Bonzon, U. Endriss, N. Maudet and J. Rossit, Rationalisation of profiles of abstract argumentation frameworks, in: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016)*, Singapore, 2016, pp. 9–13, 350–357.

[2] C. Alchourrón, P. Gärdenfors and D. Makinson, On the logic of theory change: Partial meet contraction and revision functions, *Journal of Symbolic Logic* **50**(2) (1985), 510–530. doi:10.2307/2274239.

[3] O. Arieli and M. Caminada, A QBF-based formalization of abstract argumentation semantics, *Journal of Applied Logic* **11**(2) (2013), 229–252. doi:10.1016/j.jal.2013.03.009.

[4] K. Arrow, *Social Choice and Individual Values*, 2nd edn, Yale University Press, New Haven, CT, 1963.

[5] K. Arrow, A. Sen and K. Suzumura (eds), *Handbook of Social Choice and Welfare*, Vol. 1, Elsevier Science Publishers, 2002.

[6] M. Auday, Sistemas argumentativos y agregación. Algunos resultados, in: *Epistemología e Historia de la Ciencia*, Vol. 15, Universidad Nacional de Córdoba, 2009, pp. 26–32.

[7] M. Auday, Sistemas argumentativos, unanimidad y derrota revelada, in: *Actas de las XV Jornadas de Epistemología de las Ciencias Económicas*, Facultad de Ciencias Económicas, Universidad de Buenos, Aires, 2009.

[8] E. Awad, J. Bonnefon, M. Caminada, T. Malone and I. Rahwan, Experimental assessment of aggregation rules in argumentation-enabled collective intelligence, 2016 (http://arxiv.org/abs/1604.00681).

[9] E. Awad, R. Booth, F. Tohmé and I. Rahwan, Judgement aggregation in multi-agent argumentation, *Journal of Logic and Computation* (2015), in press. doi:10.1093/logcom/exv055.

[10] P. Baroni, M. Caminada and M. Giacomin, An introduction to argumentation semantics, *The Knowledge Engineering Review* **26**(04) (2011), 365–410. doi:10.1017/S0269888911000166.

[11] G. Bodanza, Racionalidad colectiva en la argumentación social. Imposibilidad general y posibilidad restringida, in: *Filosofía e Historia de la Ciencia en el Cono Sur. Selección de Trabajos del IX Encuentro y XXV Jornadas de Epistemología e Historia de la Ciencia*, 2015, pp. 129–138.

[12] G. Bodanza and M. Auday, Social argument justification: Some mechanisms and conditions for their coincidence, in: *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Springer, Berlin, 2009, pp. 95–106. doi:10.1007/978-3-642-02906-6_10.

[13] E. Bonzon and N. Maudet, in: *On the Outcomes of Multiparty Persuasion*, P. McBurney, S. Parsons and I. Rahwan, eds, Springer, Berlin, Heidelberg, 2012, pp. 86–101.

[14] E. Bonzon, N. Maudet and S. Moretti, Coalitional games for abstract argumentation, in: *Proceedings of the 4th International Conference on Computational Models of Argument (COMMA 2014)*, IOS Press, 2014, pp. 161–172.

[15] R. Booth, E. Awad and I. Rahwan, Interval methods for judgment aggregation in argumentation, in: *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourteenth International Conference (KR2014)*, 2014, pp. 594–597.

[16] R. Booth, M. Caminada, M. Podlaszewski and I. Rahwan, Quantifying disagreement in argument-based reasoning, in: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems – Volume 1. International Foundation for Autonomous Agents and Multiagent Systems*, 2012, pp. 493–500.

[17] R. Booth and M. Podlaszewski, Using distances for aggregation in abstract argumentation, in: *Proceedings of the 26th Benelux Conference on Artificial Intelligence (BNAIC 2014)*, 2014, pp. 1–8.

[18] S. Bromuri and M. Morge, Multiparty argumentation game for consensual expansion, in: *ICAART 2013 – Proceedings of the 5th International Conference on Agents and Artificial Intelligence*, 15–18 February, 2013, Vol. 1, Barcelona, Spain, 2013, pp. 160–165.

[19] M. Caminada, On the issue of reinstatement in argumentation, in: *Proceedings of the 10th European Conference on Logics in Artificial Intelligence (JELIA 2006)*, Liverpool, UK, September 13–15, 2006, 2006, pp. 111–123. doi:10.1007/11853886_11.

[20] M. Caminada and G. Pigozzi, On judgment aggregation in abstract argumentation, *Autonomous Agents and Multi-Agent Systems* **22**(1) (2011), 64–102. doi:10.1007/s10458-009-9116-7.

[21] M. Caminada, G. Pigozzi and M. Podlaszewski, Manipulation in group argument evaluation, in: *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, International Foundation for Autonomous Agents and Multiagent Systems, 2011, pp. 121–126.

[22] C. Cayrol and M.C. Lagasquie-Schiex, On the acceptability of arguments in bipolar argumentation frameworks, in: *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Springer, Berlin, 2005, pp. 378–389. doi:10.1007/11518655_33.

[23] C. Cayrol and M.C. Lagasquie-Schiex, Gradual valuation for bipolar argumentation frameworks, in: *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Springer, Berlin, 2005, pp. 366–377. doi:10.1007/11518655_32.

[24] C. Cayrol and M.C. Lagasquie-Schiex, Bipolarity in argumentation graphs: Towards a better understanding, in: *Scalable Uncertainty Management*, Springer, Berlin, 2011, pp. 137–148. doi:10.1007/978-3-642-23963-2_12.

[25] 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) (2007), 730–753. doi:10.1016/j.artint.2007.04.012.

[26] S. Coste-Marquis, S. Konieczny, J.G. Mailly and P. Marquis, On the revision of argumentation systems: Minimal change of arguments statuses, in: *Proceedings of the Fourteenth International Conference on Principles of Knowledge Representation and Reasoning*, 2014, pp. 52–61.

[27] J. Delobelle, A. Haret, S. Konieczny, J.G. Mailly, J. Rossit and S. Woltran, Merging of abstract argumentation frameworks, in: *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2016, pp. 33–42.

[28] F. Dignum and G. Vreeswijk, Towards a testbed for multi-party dialogues, in: *Advances in Agent Communication*, Springer, Berlin, 2003, pp. 212–230.

[29] Y. Dimopoulos and P. Moraitis, Advances in argumentation-based negotiation, in: *Negotiation and Argumentation in Multi-Agent Systems: Fundamentals, Theories, Systems and Applications. Bentham E-Books*, F. Lopes and H. Coelho, eds, 2014, pp. 84–125.

[30] P.M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, *Artificial Intelligence* **77**(2) (1995), 321–357. doi:10.1016/0004-3702(94)00041-X.

[31] P.M. Dung, R. Kowalski and F. Toni, Assumption-based argumentation, in: *Argumentation in AI*, I. Rahwan and G. Simari, eds, Springer, 2009, pp. 199–218.

[32] P. Dunne, A. Hunter, P. McBurney, S. Parsons and M. Wooldridge, Weighted argument systems: Basic definitions, algorithms, and complexity results, *Artificial Intelligence* **175**(2) (2011), 457–486. doi:10.1016/j.artint.2010.09.005.

[33] P. Dunne, P. Marquis and M. Wooldridge, Argument aggregation: Basic axioms and complexity results, in: *Proceedings of the 4th International Conference on Computational Models of Argument (COMMA 2012)*, B. Verheij, S. Szeider and S. Woltran, eds, 2012, pp. 129–140.

[34] S. Dyrkolbotn and M. Walicki, Propositional discourse logic, *Synthese* **191**(5) (2014), 863–899. doi:10.1007/s11229-013-0297-x.

[35] U. Endriss and U. Grandi, Collective rationality in graph aggregation, in: *ECAI 2014. 21st European Conference on Artificial Intelligence*, T. Schaub, G. Friedrich and B. O'Sullivan, eds, Prague, Czech Republic, August 2014, 2014, pp. 291–296.

[36] D. Gabbay, Fibring argumentation frames, *Studia Logica* **93**(2) (2009), 231–295. doi:10.1007/s11225-009-9217-y.

[37] D. Gabbay, Systems of interacting argumentation networks. IFCoLog Journal of Logics and their Applications, 2014 Jun; 1(1):131–176.

[38] D. Gabbay and O. Rodrigues, A numerical approach to the merging of argumentation networks, in: *Computational Logic in Multi-Agent Systems. 13th International Workshop (CLIMA XIII)*, M. Fisher, L. van der Torre, M. Dastani and G. Governatori, eds, Montpellier, France, August 27–28, 2012, Springer, 2012, pp. 195–212.

[39] K. Grosse, C. Chesñevar, A.G. Maguitman and E. Estevez, Empowering an e-government platform through Twitter-based arguments, *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial* **15**(50) (2012), 46–56.

[40] D. Grossi, Doing argumentation theory in modal logic, Technical Report PP-2009-24, ILLC, 2009.

[41] D. Grossi, On the logic of argumentation theory, in: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1. International Foundation for Autonomous Agents and Multiagent Systems*, 2010, pp. 409–416.

[42] D. Grossi, Argumentation in the view of modal logic, in: *Argumentation in Multi-Agent Systems*, P. McBurney, I. Rahwan and S. Parsons, eds, Vol. 6614, Springer, Berlin, Heidelberg, 2011, pp. 190–208. doi:10.1007/978-3-642-21940-5_12.

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

[44] H. Jakobovits and D. Vermeir, Dialectic semantics for argumentation frameworks, in: *Proceedings of the 7th International Conference on Artificial Intelligence and Law*, ACM, 1999, pp. 53–62.

[45] H. Katsuno and A. Mendelzon, Propositional knowledge base revision and minimal change, *Artificial Intelligence* **52** (1991), 263–294. doi:10.1016/0004-3702(91)90069-V.

[46] S. Konieczny and R.P. Pérez, Propositional belief base merging or how to merge beliefs/goals coming from several sources and some links with social choice theory, *European Journal of Operational Research* **160**(3) (2005), 785–802. doi:10.1016/j.ejor.2003.06.039.

[47] D. Kontarinis, E. Bonzon, N. Maudet and P. Moraitis, Regulating multiparty persuasion with bipolar arguments: Discussion and examples, in: *Modèles Formels de L'Interaction*, Rouen, France, 2011, pp. 119–129.

[48] D. Kontarinis, E. Bonzon, N. Maudet and P. Moraitis, Picking the right expert to make a debate uncontroversial, in: *Proceedings of the 4th International Conference on Computational Models of Argument (COMMA 2012)*, IOS Press, 2012, pp. 486–497.

[49] D. Kontarinis, E. Bonzon, N. Maudet, A. Perotti, L. van der Torre and S. Villata, Rewriting rules for the computation of goal-oriented changes in an argumentation system, in: *International Workshop on Computational Logic in Multi-Agent Systems*, Springer, 2013, pp. 51–68.

[50] R. Kowalski and M. Sergot, A logic-based calculus of events, in: *Foundations of Knowledge Base Management*, Springer, Berlin, 1989, pp. 23–55. doi:10.1007/978-3-642-83397-7_2.

[51] K. Larson and I. Rahwan, Welfare properties of argumentation-based semantics, in: *Proceedings of the 2nd International Workshop on Computational Social Choice (COMSOC)*, 2008, pp. 1–12.

[52] J. Leite and J. Martins, Social abstract argumentation, in: *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, Barcelona, Catalonia, Spain, July 16–22, 2011, Vol. 11, 2011, pp. 2287–2292.

[53] N. Li, A liberal impossibility of abstract argumentation, in: *Workshop on Social Choice and Artificial Intelligence*, 2011, pp. 46–51.

[54] C. List, The theory of judgment aggregation: An introductory review, *Synthese* **187**(1) (2012), 179–207. doi:10.1007/s11229-011-0025-3.

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

[56] C. List and C. Puppe, Judgment aggregation: A survey, in: *Handbook of Rational and Social Choice*, P. Anand, P. Pattanaik and C. Puppe, eds, Oxford University Press, 2009, pp. 457–482. doi:10.1093/acprof:oso/9780199290420.003.0020.

[57] S. Coste-Marquis, C. Devred, S. Konieczny, M.C. Lagasquie-Schiex and P. Marquis, Merging argumentation systems, in: *Proceedings of the Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference*, July 9–13, Pittsburgh, Pennsylvania, USA, 2005, pp. 614–619.

[58] M. Miller and D. Osherson, Methods for distance-based judgment aggregation, *Social Choice and Welfare* **32**(4) (2009), 575–601. doi:10.1007/s00355-008-0340-x.

[59] S. Ontañón and E. Plaza, Learning and joint deliberation through argumentation in multiagent systems, in: *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*, ACM, 2007, pp. 159:1–159:8.

[60] S. Ontañón and E. Plaza, Argumentation-based information exchange in prediction markets, in: *Argumentation in Multi-Agent Systems*, I. Rahwan and P. Moraitis, eds, Springer, Berlin, 2009, pp. 181–196. doi:10.1007/978-3-642-00207-6_11.

[61] T. Pedersen and S. Dyrkolbotn, Computing consensus: A logic for reasoning about deliberative processes based on argumentation, *CoRR* (2014) (abs/1408.1647).

[62] G. Pigozzi, Belief merging and the discursive dilemma: An argument-based account to paradoxes of judgment aggregation, *Synthese* **152**(2) (2006), 285–298. doi:10.1007/s11229-006-9063-7.

[63] G. Pigozzi and L. van der Torre, Premise independence in judgment aggregation, in: *Formal Models of Belief Change in Rational Agents. No. 07351 in Dagstuhl Seminar Proceedings*, G. Bonanno, J. Delgrande, J. Lang and H. Rott, eds, Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Dagstuhl, Germany, 2007.

[64] T. Polacsek and L. Cholvy, A framework to report and to analyse a debate, in: *Proceedings of the 2011 15th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2011*, 2011, pp. 84–90.

[65] I. Rahwan and K. Larson, Mechanism design for abstract argumentation, in: *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, Estoril, Portugal, May 12–16, 2008, Vol. 2, International Foundation for Autonomous Agents and Multiagent Systems, 2008, pp. 1031–1038.

[66] I. Rahwan and K. Larson, Pareto optimality in abstract argumentation, in: *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, 2008, pp. 150–155.

[67] I. Rahwan, K. Larson and F. Tohmé, A characterization of strategy-proofness for grounded argumentation semantics, in: *IJCAI 2009, Proceedings of the 21st International Joint Conference on Artificial Intelligence*, Pasadena, California, USA, July 11–17, 2009, Citeseer, 2009, pp. 251–256.

[68] I. Rahwan and M. Podlaszewski, Complexity properties of critical sets of arguments, in: *Proceedings of the 4th International Conference on Computational Models of Argument (COMMA 2014)*, IOS Press, 2014, pp. 163–184.

[69] I. Rahwan and F. Tohmé, Collective argument evaluation as judgement aggregation, in: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1. International Foundation for Autonomous Agents and Multiagent Systems*, 2010, pp. 417–424.

[70] A. Sen, The impossibility of a Paretian liberal, *Journal of Political Economy* **78**(1) (1970), 152–157. doi:10.1086/259614.

[71] F. Tohmé, G. Bodanza and G. Simari, Aggregation of attack relations: A social-choice theoretical analysis of defeasibility criteria, in: *Foundations of Information and Knowledge Systems. 5th International Symposium (FoIKS 2008)*, Pisa, Italy, February 11–15, 2008, S. Hartmann and G. Kern-Isberner, eds, Springer, Berlin, 2008, pp. 8–23. doi:10.1007/978-3-540-77684-0_4.

[72] F. Toni and P. Torroni, Bottom-up argumentation, in: *Theorie and Applications of Formal Argumentation – First International Workshop, TAFA 2011*, Barcelona, Spain, July 16–17, 2011, 2011, pp. 249–262, Revised Selected Papers.

[73] B. Verheij, A labeling approach to the computation of credulous acceptance in argumentation, in: *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 2007, pp. 623–628.