Quantifying Disagreement in Argument-based Reasoning

Richard Booth[1], Martin Caminada[1], Mikołaj Podlaszewski[1], Iyad Rahwan[2,3]
[1]University of Luxembourg, Luxembourg
[2]Masdar Institute of Science & Technology, UAE
[3]Massachusetts Institute of Technology, USA

ABSTRACT

An argumentation framework can be seen as expressing, in an abstract way, the conflicting information of an underlying logical knowledge base. This conflicting information often allows for the presence of more than one possible reasonable position (extension/labelling) which one can take. A relevant question, therefore, is how much these positions differ from each other. In the current paper, we will examine the issue of how to define meaningful measures of distance between the (complete) labellings of a given argumentation framework. We provide concrete distance measures based on argument-wise label difference, as well as based on the notion of critical sets, and examine their properties.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence

General Terms

Algorithms, Measurement, Theory

Keywords

argumentation, distance measures, complete labellings, aggregation, belief revision

1. INTRODUCTION

Given a conflicting logical theory, an agent is faced with the problem of deciding what it could reasonably believe. As advocated in various nonmonotonic inference formalisms such as default logic [24], it is often possible to identify *multiple* reasonable positions, or so-called *extensions*. This idea has been adopted in abstract argumentation theory [14], which attempts to analyze possible extensions while abstracting away from the underlying logic. In particular, this theory views logical derivations as abstract arguments (nodes in a graph), and conflicts as defeat relations (directed arcs) over these arguments.

The presence of multiple reasonable positions raises a fundamental question: *how different are two given evaluations of a conflicting logical theory?* We attempt to answer this question in the context of abstract argumentation theory.

This question is relevant to two fundamental problems. The first problem is *argument-based belief revision*. Suppose a diplomat receives instructions to switch his position on one particular argument (see Section 3 for an example). To maintain a consistent viewpoint, the diplomat must revise his evaluation of other related arguments. Faced with multiple possibilities, the diplomat may wish to choose the one that differs the least from his initial position (e.g. to maintain credibility).

The issue of distance is also relevant to the problem of *judgement aggregation* over how a given set of arguments should be evaluated collectively by a group of agents with different opinions [11, 12, 23]. For instance it is very well possible that the members of a jury in a criminal trial all share the same information on the case (and hence have the same argumentation framework) but still have different opinions on what the verdict should be. Hence, these differences of opinion are consequences not of differences in the knowledge base but of the nature of nonmonotonic reasoning, which allows for various reasonable positions (extensions). In the context of judgement aggregation one may examine the extent to which the collective position differs from the various positions of the individual participants. Ideally, one would like to have a collective position that is closest to the collection of individual positions, for example such that the sum of its distance to each individual position is minimal.

In this paper, we examine a number of possible candidates for measuring the *distance* between different labellings (evaluations) of an argumentation graph. The paper advances the state-of-the-art in argument-based reasoning in three ways: (1) We provide the first systematic investigation of quantifying the distance between two evaluations of an argument graph; (2) We examine a number of intuitive measures and show that they fail to satisfy basic desirable postulates; (3) we come up with a measure that satisfies them all. In addition to providing many answers, our paper also raises many interesting questions to the community at the intersection between argumentation and social choice.

2. ABSTRACT ARGUMENTATION

In this section, we briefly restate some preliminaries regarding argumentation theory. For simplicity, we only consider finite argumentation frameworks.

Definition 1. An *argumentation framework* (AF for short) is a pair $\mathcal{A} = (Ar, \rightharpoonup)$, where Ar is a finite set of arguments

Appears in: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Conitzer, Winikoff, Padgham, and van der Hoek (eds.), 4-8 June 2012, Valencia, Spain.

and $\rightharpoonup \subseteq Ar \times Ar$.

We say that argument A *attacks* argument B iff $(A, B) \in \rightharpoonup$. An AF can be represented as a directed graph in which the arguments are represented as nodes and the attack relation is represented as arrows.

In the current paper, we follow the approach of [6, 10] in which the semantics of abstract argumentation is expressed in terms of *argument labellings*. The idea is to distinguish between the arguments that one accepts (that are labelled `in`), the arguments that one rejects (that are labelled `out`) and the arguments which one abstains from having an opinion about (that are labelled `undec` for "undecided").

Definition 2. Given an AF $\mathcal{A} = (Ar, \rightharpoonup)$, a labelling for $\mathcal{A}$ is a function $\mathcal{L} : Ar \rightarrow \{\texttt{in}, \texttt{out}, \texttt{undec}\}$.

Since a labelling is a function, it can be represented as a set of pairs, each consisting of an argument and a label (`in`, `out`, or `undec`). We are now ready to state the concept of *complete labelling* [6, 10].

Definition 3. Let $\mathcal{L}$ be a labelling for AF $\mathcal{A} = (Ar, \rightharpoonup)$. $\mathcal{L}$ is a complete labelling (over $\mathcal{A}$) iff for each $A \in Ar$ it holds that:

1. $\mathcal{L}(A) = \texttt{in}$ iff $\forall B \in Ar : (B \rightharpoonup A \supset \mathcal{L}(B) = \texttt{out})$
2. $\mathcal{L}(A) = \texttt{out}$ iff $\exists B \in Ar : (B \rightharpoonup A \wedge \mathcal{L}(B) = \texttt{in})$.

We denote the set of all complete labellings of $\mathcal{A}$ by $Comp_{\mathcal{A}}$.

As stated in [6, 10], complete labellings coincide with complete extensions in the sense of [14]. Moreover, the relationship between them is one-to-one. In essence, a complete extension is simply the `in`-labelled part of a complete labelling [6, 10].

The labelling approach has also been defined for other semantics, such as grounded, preferred, stable and semi-stable semantics, as well as for ideal semantics (see the overview article [2] for details). In this paper, however, we will focus on the case of complete semantics and the associated complete labellings, not only because of their relative simplicity, but also because complete labellings serve as the basis for defining labellings for various other semantics [10]. That is, semantics like grounded [14], preferred [14], stable [14], semi-stable [9], ideal [15] and eager [7] in essence select subsets of the set of all complete labellings (see [2]). Since the approach in the current paper is to compare any arbitrary pair of complete labellings, our results are directly applicable also to the aforementioned semantics.[1]

Example 1. Consider a simple argumentation framework $\mathcal{A} = (Ar, \rightharpoonup)$ with $Ar = \{A, B, C\}$ and $\rightharpoonup = \{(A, B), (B, A), (B, C)\}$. Then $Comp_{\mathcal{A}} = \{\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3\}$, where each $\mathcal{L}_i$ may be visualised in Fig. 1. In this and subsequent diagrams, a node with a solid line indicates an `in` label, a dotted line indicates `out` and a grey node indicates `undec`. Thus for example the first labelling $\mathcal{L}_1 = \{(A, \texttt{in}), (B, \texttt{out}), (C, \texttt{in})\}$.

[1]Another point to mention is that it has been proved that complete-based semantics (that is, semantics whose sets of extensions/labellings are subsets of the set of all complete extensions/labellings), when used for the purpose of logical inference, tend to produce fully instantiated argumentation formalisms that satisfy reasonable properties in the sense of [8, 22].

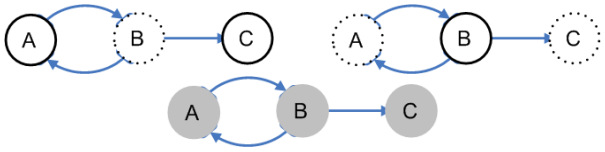

Figure 1: Three possible complete labellings $\mathcal{L}_1, \mathcal{L}_2$ and $\mathcal{L}_3$

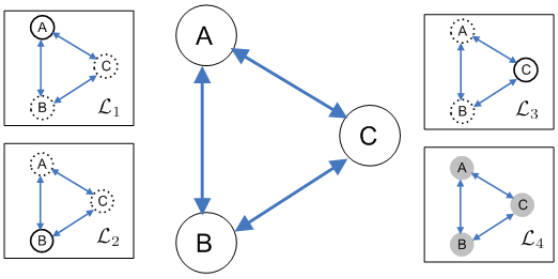

Figure 2: The AF $\mathcal{A}_1$ and its four possible complete labellings $\mathcal{L}_1$-$\mathcal{L}_4$

3. BELIEF REVISION AND JUDGEMENT AGGREGATION

Consider three arguments about global warming, each one grounded in some scientific evidence, with the following associated conclusions:

- A: Global warming is mainly caused by volcanic activity.
- B: Global warming is mainly caused by natural variation in solar radiation.
- C: Global warming is a human-induced phenomenon.

Clearly, it is not possible to subscribe to both arguments A and B, since they attribute global warming to different major causes. However, both these arguments attack argument C, which attributes global warming to human activity. This situation can be modelled with the AF $\mathcal{A}_1$ shown in Fig.2. Here there are four possible complete labellings $\mathcal{L}_1$-$\mathcal{L}_4$, which are also depicted.

Suppose a diplomat was initially adopting the position corresponding to labelling $\mathcal{L}_3$ in Fig. 2, which focusses on human activity (argument C) as the main cause of global warming.

However, recent elections in his home country gave rise to a new climate-sceptical government, which requires the diplomat to change his position in order to no longer accept argument C (that is: to revise his complete labelling to one in which C is no longer labelled `in`). He is now faced with two possible revisions of his original position $\mathcal{L}_3$. On one hand, he can switch to labelling $\mathcal{L}_1$ or $\mathcal{L}_2$, citing the alternative theory. On the other hand, he could switch to labelling $\mathcal{L}_4$, admitting that the matter cannot be decided.

Let us assume that politicians like to maintain a reputation of being generally consistent. Therefore, when they switch their points of view, they like to minimize the extent to which they deviate from their original positions. In the

above example, it is far from obvious which revision of the original position $\mathcal{L}_3$ is less dramatic. On one hand, switching to the alternative theories $\mathcal{L}_1$ or $\mathcal{L}_2$ keeps the status of at least one argument the same, while switching to labelling $\mathcal{L}_4$ requires changing the status of all arguments involved. On the other hand, switching the status of argument C from being fully accepted to being completely rejected (as in $\mathcal{L}_1$ or $\mathcal{L}_2$) seems more severe than simply moving to a position of indecision (as in $\mathcal{L}_4$).

Following up on the example above, suppose we have a panel consisting of three scientists, with two supporting position $\mathcal{L}_1$, and one supporting position $\mathcal{L}_2$. Suppose the scientists want to reach a collective position that is closest to their respective individual positions, in order to minimize the degree to which they individually deviate from their original positions. To achieve this, should all of them concede to the third undecided position $\mathcal{L}_4$, as is suggested in [11]? Or should the third scientist individually concede to position $\mathcal{L}_1$, ensuring the first two stick to their view? The answer to this question relies crucially on how we quantify the distance between the different positions.

The examples above highlight the need for a systematic approach to identifying the extent to which two positions differ, ideally creating a reliable quantitative measure of distance between different complete labellings.

4. DISTANCE BETWEEN LABELLINGS

The problem we are interested is the following:

> Given an AF $\mathcal{A}$, and given two complete labellings $\mathcal{S}$ (the *source* labelling) and $\mathcal{T}$ (the *target* labelling) over $\mathcal{A}$, how can we quantify the *distance* from $\mathcal{S}$ to $\mathcal{T}$, denoted $d(\mathcal{S}, \mathcal{T})$?

Of course we don't just want a method which applies to only one AF, we want a method to be able to do this for *any* given $\mathcal{A}$.

Definition 4. A *labelling distance* (for AF $\mathcal{A}$) is a function $d : \mathit{Comp}_{\mathcal{A}} \times \mathit{Comp}_{\mathcal{A}} \to \mathbb{N}$. A *labelling distance method* is a function which assigns to every AF $\mathcal{A}$ a labelling distance for $\mathcal{A}$.

In the following sections we will provide a few concrete definitions of distance functions. But first, are there any properties which we should expect such a function to satisfy?

4.1 Properties for distance methods

In mathematics, when formalising the notion of distance it is common to require that d be a *metric*. In our present setting that means that the following hold for all complete labellings $\mathcal{S}, \mathcal{T}, \mathcal{U}$ over a given AF $\mathcal{A}$:

(dm1) $d(\mathcal{S}, \mathcal{S}) = 0$
(dm2) $d(\mathcal{S}, \mathcal{T}) > 0$ if $\mathcal{S} \neq \mathcal{T}$
(dm3) $d(\mathcal{S}, \mathcal{T}) = d(\mathcal{T}, \mathcal{S})$ (Symmetry)
(dm4) $d(\mathcal{S}, \mathcal{T}) \leq d(\mathcal{S}, \mathcal{U}) + d(\mathcal{U}, \mathcal{T})$ (Triangle inequality)

Also, let's define the following binary relation over $\mathit{Comp}_{\mathcal{A}}$, given a fixed source complete labelling $\mathcal{S}$:

$$\mathcal{T}_1 \leq_{\mathcal{S}} \mathcal{T}_2 \text{ iff } \forall A \, (\mathcal{T}_1(A) = \mathcal{S}(A) \vee \mathcal{T}_2(A) = \mathcal{T}_1(A))$$

$\mathcal{T}_1 \leq_{\mathcal{S}} \mathcal{T}_2$ means that every argument that $\mathcal{T}_1$ labels differently from $\mathcal{S}$, is labelled equally differently by $\mathcal{T}_2$. Thus $\mathcal{T}_2$ differs from $\mathcal{S}$ at least as much as $\mathcal{T}_1$ does. It can be shown that $\leq_{\mathcal{S}}$ is a partial order over $\mathit{Comp}_{\mathcal{A}}$ with minimum element $\mathcal{S}$, i.e., $\mathcal{S} \leq_{\mathcal{S}} \mathcal{T}$ for all $\mathcal{T} \in \mathit{Comp}_{\mathcal{A}}$. Let $<_{\mathcal{S}}$ denote the strict version of $\leq_{\mathcal{S}}$, i.e., $\mathcal{T}_1 <_{\mathcal{S}} \mathcal{T}_2$ iff both $\mathcal{T}_1 \leq_{\mathcal{S}} \mathcal{T}_2$ and $\mathcal{T}_2 \not\leq_{\mathcal{S}} \mathcal{T}_1$. Thus the following might seem to be a reasonable requirement on a distance function d:

(dm5) If $\mathcal{T}_1 <_{\mathcal{S}} \mathcal{T}_2$ then $d(\mathcal{S}, \mathcal{T}_1) < d(\mathcal{S}, \mathcal{T}_2)$ (Disagreement monotonicity)

To see why this might be reasonable, note that $\mathcal{T}_1 <_{\mathcal{S}} \mathcal{T}_2$ means that for every argument on which $\mathcal{T}_1$ disagrees with $\mathcal{S}$, the labelling $\mathcal{T}_2$ disagrees with $\mathcal{S}$ in exactly the same way, but that there exists at least one argument on which $\mathcal{T}_2$ disagrees with $\mathcal{S}$, but for which $\mathcal{T}_1$ and $\mathcal{S}$ agree. In this case it seems as though $\mathcal{T}_2$ is making strictly more changes to $\mathcal{S}$ than $\mathcal{T}_1$ is, and so d should also endorse this conclusion. It is not difficult to show that if d satisfies both **(dm1)** and **(dm5)** then it satisfies **(dm2)**.

We can also describe a postulate which is stronger than **(dm5)**. To express this property we first define the following ordering over $\mathit{Comp}_{\mathcal{A}}$, given any source labelling $\mathcal{S}$ and target labellings $\mathcal{T}_1, \mathcal{T}_2$:

$$\mathcal{T}_1 \leq^b_{\mathcal{S}} \mathcal{T}_2 \text{ iff } \forall A \left(\begin{array}{l} \mathcal{T}_1(A) = \mathcal{S}(A) \vee \mathcal{T}_1(A) = \mathcal{T}_2(A) \\ \vee [\mathcal{T}_1(A) = \texttt{undec} \wedge \mathcal{S}(A) \neq \mathcal{T}_2(A)] \end{array} \right)$$

Like $\leq_{\mathcal{S}}$, the ordering $\leq^b_{\mathcal{S}}$ forms a partial order with minimum element $\mathcal{S}$. The superscript "b" on $\leq^b_{\mathcal{S}}$ may be thought of as standing for "*between*", since $\mathcal{T}_1 \leq^b_{\mathcal{S}} \mathcal{T}_2$ is merely expressing that, for all $A \in Ar$, $\mathcal{T}_1(A)$ lies on a path *between* $\mathcal{S}(A)$ and $\mathcal{T}_2(A)$, assuming the neighbourhood graph $\texttt{in} - \texttt{undec} - \texttt{out}$ over the labels. We may then propose the following:

(dm5+) If $\mathcal{T}_1 <^b_{\mathcal{S}} \mathcal{T}_2$ then $d(\mathcal{S}, \mathcal{T}_1) < d(\mathcal{S}, \mathcal{T}_2)$ (Betweenness monotonicity)

where $<^b_{\mathcal{S}}$ is the strict part of the relation $\leq^b_{\mathcal{S}}$. Since clearly $\mathcal{T}_1 <_{\mathcal{S}} \mathcal{T}_2$ implies $\mathcal{T}_1 <^b_{\mathcal{S}} \mathcal{T}_2$ we have that **(dm5+)** is indeed a strengthening of **(dm5)**.

5. SUM-BASED DISTANCE

Our first family of distance functions is about simply finding the raw quantity of disagreement between two complete labellings. We can do this in terms of difference between labels, that is, we assume we have some measure of disagreement $\mathit{diff}(\mathtt{x}, \mathtt{y})$ for any $\mathtt{x}, \mathtt{y} \in \{\mathtt{in}, \mathtt{out}, \mathtt{undec}\}$ between the different labels, and then obtain the distance between two labellings by summing the differences between *all* arguments in the AF $(Ar, \rightharpoonup)$ under consideration, i.e., take

$$d(\mathcal{S}, \mathcal{T}) = \sum_{A \in Ar} \mathit{diff}(\mathcal{S}(A), \mathcal{T}(A)). \quad (1)$$

Definition 5. If the function d can be defined from some function $\mathit{diff} : \{\mathtt{in}, \mathtt{out}, \mathtt{undec}\}^2 \to \mathbb{N}$ as in (1) then we say d is a *simple diff-based distance method*.

It turns out that the results in this paper depend only on a few fundamental requirements on *diff*, encapsulated in the following definition:

Definition 6. A *basic label difference measure* is a function $\mathit{diff} : \{\mathtt{in}, \mathtt{out}, \mathtt{undec}\}^2 \to \mathbb{N}$ which satisfies the following properties, for all $\mathtt{x}, \mathtt{y} \in \{\mathtt{in}, \mathtt{out}, \mathtt{undec}\}$:

(diff 1) $\mathit{diff}(\mathtt{x},\mathtt{x}) = 0$
(diff 2) $\mathit{diff}(\mathtt{x},\mathtt{y}) = \mathit{diff}(\mathtt{y},\mathtt{x})$
(diff 3) $\mathit{diff}(\mathtt{in},\mathtt{out}) > 0$
(diff 4) $\mathit{diff}(\mathtt{in},\mathtt{undec}) = \mathit{diff}(\mathtt{out},\mathtt{undec})$

Note that this means, in effect, any simple diff-based measure based on a basic label difference measure is completely specified by 2 quantities: $\mathit{diff}(\mathtt{in},\mathtt{undec})$ and $\mathit{diff}(\mathtt{in},\mathtt{out})$, which may respectively be thought of as the costs attached to a *soft* and *hard* conflict. From now on any unspecified *diff*-measure will be assumed to satisfy **(diff 1)**-**(diff 4)**.

PROPOSITION 1. *If d is a simple diff-based distance method defined via a basic label difference measure then d satisfies* **(dm1)** *and* **(dm3)**.

(We remark that Propositions 1-4 in this section actually all follow as corollaries of a more general result, Theorem 1, in Section 6.1). As we will see below, the remaining distance properties from the previous section can easily be captured by placing further, optional, constraints on *diff*. Let us take in a few concrete examples.

Measuring incompatibility

The property **(diff 3)** ensures that a hard conflict always contributes a strictly positive value. But note we do not require a soft conflict to do the same. That is, we do not insist on the following strengthening of **(diff 3)**:

(diff 3+) $\mathit{diff}(\mathtt{x},\mathtt{y}) > 0$ for $\mathtt{x} \neq \mathtt{y}$

In this way we allow *diff*-measures such as the following, which is inspired by the work of [11]. A labelling $\mathcal{L}_1$ is *compatible* with labelling $\mathcal{L}_2$ (written as $\mathcal{L}_1 \approx \mathcal{L}_2$) iff there is no argument A such that either $[\mathcal{L}_1(A) = \mathtt{in}$ and $\mathcal{L}_2(A) = \mathtt{out}]$ or $[\mathcal{L}_1(A) = \mathtt{out}$ and $\mathcal{L}_2(A) = \mathtt{in}]$. The idea behind compatibility is to give a rough impression of how difficult it is to publicly defend a position (labelling) that is not one's own. Although it might be possible to publicly accept or reject an argument which one privately has no opinion about (undec), or to remain silent about an argument that one privately accepts or rejects, it is significantly more difficult to publicly accept an argument which one privately rejects (and vice versa). Our first concrete measure of distance makes the distance zero if the two labellings are compatible, and measures the "degree of incompatibility" if they are not.

$$\mathit{diff}^{\approx}(\mathtt{in},\mathtt{out}) = 1, \qquad \mathit{diff}^{\approx}(\mathtt{in},\mathtt{undec}) = 0.$$

This leads to a function $d^{\approx}$ (defined using $\mathit{diff}^{\approx}$ via (1)) which is more like a "measure of conflict" between $\mathcal{S}$ and $\mathcal{T}$. Measure $d^{\approx}$ fails to satisfy **(dm2)**, as can be seen in Fig. 1, where $d^{\approx}(\mathcal{L}_1,\mathcal{L}_3) = 0$. If we *do* insist on **(diff 3+)** then we ensure not only **(dm2)** but also **(dm5)**:

PROPOSITION 2. *If d is a simple diff-based distance method defined via a basic label difference measure which satisfies* **(diff 3+)** *then d satisfies* **(dm5)** *(and hence also* **(dm2)***).*

Hamming distance

A very simple example of a *diff*-measure satisfying **(diff 3+)** is as follows:

$$\mathit{diff}^{H}(\mathtt{in},\mathtt{out}) = \mathit{diff}^{H}(\mathtt{in},\mathtt{undec}) = 1.$$

Then the distance between $\mathcal{S}$ and $\mathcal{T}$ boils down to the number of arguments on which $\mathcal{S}$ and $\mathcal{T}$ differ, i.e., the Hamming distance between $\mathcal{S}$ and $\mathcal{T}$. Let d^H denote the distance defined using diff^H. Consider for instance the results for Fig. 1, where we see that $d^H(\mathcal{L}_1,\mathcal{L}_2) = 3 = d^H(\mathcal{L}_1,\mathcal{L}_3)$. Thus, according to d^H, labellings $\mathcal{L}_2$ and $\mathcal{L}_3$ are equidistant from $\mathcal{L}_1$. However it might be thought that the change between $\mathcal{L}_1$ and $\mathcal{L}_2$ is more "drastic" than that between $\mathcal{L}_1$ and $\mathcal{L}_3$, since it involves a complete swing in the status of its arguments from in (resp. out) to out (resp. in). This example demonstrates that d^H fails to satisfy **(dm5+)**, since $\mathcal{L}_3 <^b_{\mathcal{L}_1} \mathcal{L}_2$ but $d^H(\mathcal{L}_1,\mathcal{L}_3) \not< d^H(\mathcal{L}_1,\mathcal{L}_2)$. Shouldn't the difference between in and out be strictly greater than the difference between in (or out) to undec? In other words we might expect:

(diff 5) $\mathit{diff}(\mathtt{in},\mathtt{out}) > \mathit{diff}(\mathtt{in},\mathtt{undec})$

PROPOSITION 3. *If d is a simple diff-based distance method defined via a basic label difference measure which satisfies* **(diff 3+)** *and* **(diff 5)** *then d satisfies* **(dm5+)**.

Refined Hamming distance

An easy way to define a basic label difference measure which satisfies both **(diff 3+)** and **(diff 5)** is to set:

$$\mathit{diff}^{rh}(\mathtt{in},\mathtt{out}) = 2, \qquad \mathit{diff}^{rh}(\mathtt{in},\mathtt{undec}) = 1,$$

where rh stands for "refined Hamming". Note $\mathit{diff}^{rh}(\mathtt{x},\mathtt{y})$ may be thought of as the length of the shortest path between x and y in the neighbourhood graph $\mathtt{in} - \mathtt{undec} - \mathtt{out}$ over the labels. We denote by d^{rh} the distance obtained by plugging diff^{rh} into (1). Going back to Fig. 1, we have $d^{rh}(\mathcal{L}_1,\mathcal{L}_2) = 3 \times \mathit{diff}^{rh}(\mathtt{in},\mathtt{out}) = 6$ and $d^{rh}(\mathcal{L}_1,\mathcal{L}_3) = 3 \times \mathit{diff}^{rh}(\mathtt{in},\mathtt{undec}) = 3$, yielding the expected $d^{rh}(\mathcal{L}_1,\mathcal{L}_3) < d^{rh}(\mathcal{L}_1,\mathcal{L}_3)$. Propositions 1-3 aready tell us d^{rh} satisfies all the distance properties from the previous section. The only one which remains is the triangle inequality **(dm4)**. But in fact this too is satisfied, owing to the fact that diff^{rh} satisfies the following (which implies **(diff 3+)** for basic label difference measures):

(diff 3++) $2 \times \mathit{diff}(\mathtt{in},\mathtt{undec}) \geq \mathit{diff}(\mathtt{in},\mathtt{out})$

This property actually ensures that *diff* itself satisfies the triangle inequality over the set of labels.

PROPOSITION 4. *If d is a simple diff-based distance method defined via a basic label difference measure which satisfies* **(diff 3++)** *then d satisfies* **(dm4)**.

Since diff^H obviously satisfies **(diff 3++)** this means we also get that d^H satisfies **(dm4)**. The incompatibility distance $d^{\approx}$, however, does not satisfy **(dm4)**, as can be seen in Fig. 1, where $d^{\approx}(\mathcal{L}_1,\mathcal{L}_2) = 3 > 0 = d^{\approx}(\mathcal{L}_1,\mathcal{L}_3) + d^{\approx}(\mathcal{L}_3,\mathcal{L}_2)$.

6. CRITICAL SETS APPROACHES

Suppose we have the complete labelling $\mathcal{S}$ shown at the top of Fig. 3 over an AF containing eight arguments $\{A, B, C, D, E, F, G, H\}$. As usual a node with a solid line denotes the argument is in, while a dotted line denotes out. Now consider the two target labellings $\mathcal{T}_1$ and $\mathcal{T}_2$ shown below it. $\mathcal{T}_1$ is obtained from $\mathcal{S}$ by leaving the labels of A, B, C, D as they are and inverting the labels of the four arguments E, F, G, H. For $\mathcal{T}_2$ we leave E, F, G, H untouched and invert the labels of the four arguments A, B, C, D. The question is: which of $\mathcal{T}_1$, $\mathcal{T}_2$ is closer to $\mathcal{S}$? Or are they both equally close?

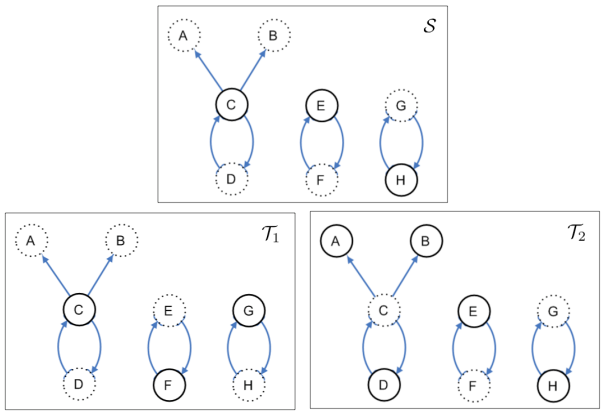

Figure 3: Source labelling $\mathcal{S}$ and 2 target labellings $\mathcal{T}_1$, $\mathcal{T}_2$

Let's consider what a simple diff-based distance function d has to say about this. One can see that we will get $d(\mathcal{S},\mathcal{T}_1) = 4 \times \mathit{diff}(\texttt{in},\texttt{out}) = d(\mathcal{S},\mathcal{T}_2)$. Thus *any* simple diff-based distance will judge $\mathcal{T}_1$ and $\mathcal{T}_2$ as equidistant from $\mathcal{S}$.

However, on reflection it seems we can be more reasonable and say that $\mathcal{T}_2$ is closer to $\mathcal{S}$. Intuitively the reason is based on the observation that disagreement between $\mathcal{S}$ and $\mathcal{T}_2$ involves a higher degree of "contagion". If two agents only differ in their opinions on argument C (or only on D), this would suffice to determine their disagreement over all other arguments in that connected component (namely A, B, C, and D). On the other hand, when comparing $\mathcal{S}$ with $\mathcal{T}_2$, two agents would have to at least disagree (fundamentally let's say) on two arguments in order for this emerge.[2]

How can we make this intuition precise? We now investigate two possible ways in which the simple diff-based approach can be refined in order to take this into account. We will see that the first one, although intuitive, is flawed.

6.1 Critical subsets approach

The first idea comes from a concept introduced by Gabbay [16]. Instead of looking at all arguments, one specifically focuses on the *critical subsets*.

Definition 7. Given an AF $\mathcal{A} = (Ar, \rightharpoonup)$, a subset $X \subseteq Ar$ is *critical* iff for any $\mathcal{L}_1$, $\mathcal{L}_2 \in \mathit{Comp}_{\mathcal{A}}$ we get $\mathcal{L}_1 = \mathcal{L}_2$ whenever $\mathcal{L}_1$ and $\mathcal{L}_2$ agree on the arguments in X. We denote the set of critical subsets for $\mathcal{A}$ by $\mathit{crit}(\mathcal{A})$.

In other words a critical subset for $\mathcal{A}$ is a set of arguments whose status is enough to determine the status of *all* the arguments in Ar. Clearly at least one critical subset will always exist, for Ar is obviously critical. We are interested in the *minimal* critical subsets.

[2]A similar intuition to this can be found in [5] in the context of reasoning about action and belief update. The idea there is that there might exist some *causal* links between the value of one literal and that of another, which should be taken into account when calculating how much one possible world, i.e., conjunction of literals, differs from another. If the change in value of one literal is caused by another, then this change should not count towards calculating the difference.

Definition 8. We denote the collection of set-theoretically minimal subsets of $\mathit{crit}(\mathcal{A})$ by $\mathit{mincrit}(\mathcal{A})$, i.e.,
$\mathit{mincrit}(\mathcal{A}) \stackrel{\text{def}}{=} \{X \in \mathit{crit}(\mathcal{A}) \mid \nexists Y (Y \in \mathit{crit}(\mathcal{A}) \wedge Y \subset X)\}$.

If we look at the AF of Fig. 3 one can check that one critical subset is $X_1 = \{C, E, G\}$, since, the label of E (respectively G) determines the label of F (respectively H), while the label of C determines the labels of A, B and D. Indeed if C is `in` then A, B and D must all be `out`, if C is `out` then A, B, D must all be `in`, while if C is `undec` then A, B, D must all be `undec` too.

So, the first idea would be, given a basic label difference measure diff, to pick some minimal critical subset X and then just define, for all $\mathcal{S}, \mathcal{T} \in \mathit{Comp}_{\mathcal{A}}$, $d'(\mathcal{S},\mathcal{T}) = d_X(\mathcal{S},\mathcal{T})$, where

$$d_X(\mathcal{S},\mathcal{T}) \stackrel{\text{def}}{=} \sum_{A \in X} \mathit{diff}(\mathcal{S}(A), \mathcal{T}(A)). \tag{2}$$

Formally, the *critical sets distance method cd* is defined via a function $\mathbb{C}$ which selects for each $\mathcal{A}$ an element of $\mathit{mincrit}(\mathcal{A})$ and then sets $\mathit{cd}(\mathcal{S},\mathcal{T}) = d_{\mathbb{C}(\mathcal{A})}(\mathcal{S},\mathcal{T})$ for any $\mathcal{S}, \mathcal{T} \in \mathit{Comp}_{\mathcal{A}}$.

Example 2. Taking the complete labellings $\mathcal{S}$ and $\mathcal{T}_1, \mathcal{T}_2$ in Fig. 3, and taking $\mathbb{C}(\mathcal{A}) = \{C, E, G\}$ we get $\mathit{cd}(\mathcal{S},\mathcal{T}_1) = 2 \times \mathit{diff}(\texttt{in},\texttt{out})$ and $\mathit{cd}(\mathcal{S},\mathcal{T}_2) = \mathit{diff}(\texttt{in},\texttt{out})$. Thus $\mathcal{T}_2$ is deemed closer to $\mathcal{S}$ than $\mathcal{T}_1$ is.

The distance function d_X in (2) actually fares rather well when measured against the properties for distance functions from earlier, provided diff is sufficiently well-behaved:

THEOREM 1. *Let $X \in \mathit{crit}(\mathcal{A})$ and let d_X be defined from diff as in (2). Then d_X satisfies* **(dm1)** *and* **(dm3)**. *Furthermore:*
(i). *If diff satisfies* **(diff 3+)** *then d_X satisfies* **(dm5)** *(and hence also* **(dm2)***).*
(ii). *If diff satisfies* **(diff 3+)** *and* **(diff 5)** *then d_X satisfies* **(dm5+)**.
(iii). *If diff satisfies* **(diff 3++)** *then d_X satisfies* **(dm4)**.

PROOF. (*Outline*) **(dm1)** and **(dm3)** follow immediately from **(diff 1)** and **(diff 3)** respectively.
(i). First it is easy to check that if $\mathcal{T}_1 \leq_{\mathcal{S}} \mathcal{T}_2$ then, for all $A \in X$ (in fact for all $A \in Ar$),

$$\mathit{diff}(\mathcal{S}(A), \mathcal{T}_1(A)) \leq \mathit{diff}(\mathcal{S}(A), \mathcal{T}_2(A)) \tag{3}$$

(since either the left-hand side equals 0 or both sides are equal). If moreover $\mathcal{T}_1 <_{\mathcal{S}} \mathcal{T}_2$ then $\mathcal{T}_1 \neq \mathcal{T}_2$ and so, since X is critical, there exists $A^* \in X$ such that $\mathcal{T}_1(A^*) \neq \mathcal{T}_2(A^*)$. From this and $\mathcal{T}_1 \leq_{\mathcal{S}} \mathcal{T}_2$ we know $\mathcal{T}_1(A^*) = \mathcal{S}(A^*)$, hence $\mathit{diff}(\mathcal{S}(A^*), \mathcal{T}_1(A^*)) = 0 < \mathit{diff}(\mathcal{S}(A^*), \mathcal{T}_2(A^*))$ (the last inequality following from **(diff 3+)**). Hence the inequality (3) is strict for at least one argument in X and thus $d_X(\mathcal{S},\mathcal{T}_1) < d_X(\mathcal{S},\mathcal{T}_2)$.
(ii). If $\mathcal{T}_1 \leq^b_{\mathcal{S}} \mathcal{T}_2$ then, for all $A \in X$ (in fact all $A \in Ar$), either *(a)* $\mathcal{T}_1(A) = \mathcal{S}(A)$, or *(b)* $\mathcal{T}_1(A) = \mathcal{T}_2(A)$, or *(c)* $\mathcal{T}_1(A) = \texttt{undec}$ and [$(\mathcal{S}(A) = \texttt{in}$ and $\mathcal{T}_2(A) = \texttt{out})$ or vice versa]. In cases *(a)*, *(b)* inequality (3) holds as in part *(i)* above, while in *(c)* we get a strict inequality due to **(diff 5)**. If $\mathcal{T}_1 <^b_{\mathcal{S}} \mathcal{T}_2$ then $\mathcal{T}_1 \neq \mathcal{T}_2$ so, since X is critical there is some $A^* \in X$ such that $\mathcal{T}_1(A^*) \neq \mathcal{T}_2(A^*)$. Then either we are in the same situation as in *(i)* above, or case *(c)* obtains. Either way the inequality (3) will be strict for A^* and so $d_X(\mathcal{S},\mathcal{T}_1) < d_X(\mathcal{S},\mathcal{T}_2)$.

(iii). Follows from the fact that **(diff 3++)** ensures *diff* itself satisfies the triangle inequality, which lifts straightforwardly to d_X. □

Note the above result holds taking X to be *any* critical subset, not only the minimal ones. By taking $X = Ar$ we thus obtain Propositions 1-4 from Section 5 as corollaries.

One problem is that more than one minimal critical subset may exist. For example in the above example one can check that another minimal critical subset can be obtained by exchanging A for D to obtain $X_2 = \{A, E, G\}$. Indeed one can exchange any argument in the leftmost component. One could also replace E by F or G by H. We would like the distance (or at least the similarity ordering induced by it) to be independent of the particular minimal critical subset we use. Is it possible that we might get $d_{X_1}(\mathcal{S}, \mathcal{T}_1) \neq d_{X_2}(\mathcal{S}, \mathcal{T}_2)$ for different minimal critical subsets X_1, X_2? In the above example the answer is no, but unfortunately this does not always hold in general, as the next example shows.

Example 3. Let us return to the AF $\mathcal{A}_1$ depicted in Fig. 2. It is not the case that by knowing the label of one argument we know the full complete labelling, however, one can check that if we know the label of any *pair* of arguments, we automatically know the label of the third. Thus we have $mincrit(\mathcal{A}_1) = \big\{\{A, B\}, \{A, C\}, \{B, C\}\big\}$. We have $d_{\{A,B\}}(\mathcal{L}_1, \mathcal{L}_2) = 2 \times diff(\mathtt{in}, \mathtt{out})$ and $d_{\{A,B\}}(\mathcal{L}_1, \mathcal{L}_3) = diff(\mathtt{in}, \mathtt{out})$. Thus if we focus on the critical subset $\{A, B\}$ we obtain that $\mathcal{L}_3$ is closer to $\mathcal{L}_1$ than $\mathcal{L}_2$ is. But if instead we focus on critical subset $\{A, C\}$ we obtain the opposite conclusion, for $d_{\{A,C\}}(\mathcal{L}_1, \mathcal{L}_2) = diff(\mathtt{in}, \mathtt{out})$ and $d_{\{A,C\}}(\mathcal{L}_1, \mathcal{L}_3) = 2 \times diff(\mathtt{in}, \mathtt{out})$.

This sensitivity to the choice of critical subset is somewhat undesirable. Furthermore, as we will see next, even though *cd* can easily be made to satisfy all the distance properties we have presented thus far, there are some other, highly intuitive, postulates that it fails to validate.

6.2 Symmetry properties

The next distance properties we propose come from symmetry considerations. The idea is that applying the distance measure over AFs which are in some sense *equivalent* should yield equivalent results. In the context of argumentation semantics, such a property has been referred to as the language independence principle [3]. We are interested in describing a similar property in the context of distance measures. We begin with the common idea of graph-isomorphism, applied to argumentation frameworks.

Definition 9. Let $\mathcal{A}_1 = (Ar_1, \rightharpoonup_1)$ and $\mathcal{A}_2 = (Ar_2, \rightharpoonup_2)$ be two AFs. An *isomorphism from $\mathcal{A}_1$ to $\mathcal{A}_2$* is any bijection $g : Ar_1 \to Ar_2$ such that, for all $A, B \in Ar_1$, $A \rightharpoonup_1 B$ iff $g(A) \rightharpoonup_2 g(B)$. In the special case when $\mathcal{A}_1 = \mathcal{A}_2$ we call g an *automorphism*.

So basically an isomorphism just changes the names of arguments – or in the case of automorphism permutes them – while preserving the attack structure. Of course if g is an isomorphism from $\mathcal{A}_1$ to $\mathcal{A}_2$ then g^{-1} is an isomorphism from $\mathcal{A}_2$ to $\mathcal{A}_1$.

If g is an isomorphism from $\mathcal{A}_1$ to $\mathcal{A}_2$ then we can extend g to a function which converts any labelling $\mathcal{S}$ for $\mathcal{A}_1$ into a labelling $g(\mathcal{S})$ for $\mathcal{A}_2$. We define labelling $g(\mathcal{S})$ simply by taking $[g(\mathcal{S})](A) = \mathcal{S}(g^{-1}(A))$ for all $A \in Ar_2$.

PROPOSITION 5. *Let g be an isomorphism from $\mathcal{A}_1$ to $\mathcal{A}_2$. If $\mathcal{S} \in Comp_{\mathcal{A}_1}$ then $g(\mathcal{S}) \in Comp_{\mathcal{A}_2}$.*

The following property says that the distance should be the same for isomorphic AFs. This is in line with the intuition that an argument is characterised completely by its interactions with the other arguments.

(Iso) If g is an isomorphism from $\mathcal{A}_1$ to $\mathcal{A}_2$ then $d_{\mathcal{A}_1}(\mathcal{S}, \mathcal{T}) = d_{\mathcal{A}_2}(g(\mathcal{S}), g(\mathcal{T}))$

Note that this property differs from our previous distance properties in that whereas they dealt with a *fixed* AF $\mathcal{A}$ as given, this rule relates distance between labellings over *different*, but related, argumentation frameworks. Technically speaking, while all the previous rules are properties of the *labelling distance* $d_{\mathcal{A}}$ for fixed $\mathcal{A}$, **(Iso)** is a property of the *distance method*, i.e., the mapping $\mathcal{A} \mapsto d_{\mathcal{A}}$ (Definition 4). In the case of automorphism we get the special case:

(Auto) If g is an automorphism on $\mathcal{A}$ then $d_{\mathcal{A}}(\mathcal{S}, \mathcal{T}) = d_{\mathcal{A}}(g(\mathcal{S}), g(\mathcal{T}))$

The distance measure *cd* fails even to satisfy **(Auto)**, as the following example shows:

Example 4. Consider $\mathcal{A}_1$ in Fig. 2 and consider the mapping g such that $g(A) = B$, $g(B) = C$ and $g(C) = A$. It is easy to see that g is an automorphism on $\mathcal{A}_1$. Assume $\mathbb{C}(\mathcal{A}_1) = \{A, B\}$. Recall $\mathcal{L}_1 = \{(A, \mathtt{in}), (B, \mathtt{out}), (C, \mathtt{out})\}$ and $\mathcal{L}_3 = \{(A, \mathtt{out}), (B, \mathtt{out}), (C, \mathtt{in})\}$. So $g(\mathcal{L}_1) = \{(A, \mathtt{out}), (B, \mathtt{in}), (C, \mathtt{out})\}$ and $g(\mathcal{L}_3) = \{(A, \mathtt{in}), (B, \mathtt{out}), (C, \mathtt{out})\}$. Then if **(Auto)** were satisfied we would expect $cd(\mathcal{L}_1, \mathcal{L}_3) = d_{\{A,B\}}(\mathcal{L}_1, \mathcal{L}_3) = d_{\{A,B\}}(g(\mathcal{L}_1), g(\mathcal{L}_3))$, but $d_{\{A,B\}}(\mathcal{L}_1, \mathcal{L}_3) = diff(\mathtt{in}, \mathtt{out}) \neq 2 \times diff(\mathtt{in}, \mathtt{out}) = d_{\{A,B\}}(g(\mathcal{L}_1), g(\mathcal{L}_3))$. Note this example assumes $\mathbb{C}(\mathcal{A}_1) = \{A, B\}$, but it should be clear that counterexamples can also be found if either of the other two elements of $mincrit(\mathcal{A}_1)$ were selected.

Summarising this section so far, we have managed to find a distance method *cd* which respects the intuitions of the example of Fig. 3, but at the expense of violating what seem to be a highly desirable postulates (**(Iso)** and **(Auto)**) for distance methods. Is there a distance method which can satisfy *all* our desiderata? We shall now see that the answer is yes.

6.3 Distance via issue-wise label difference

We want to capture the idea that the labels of two arguments are "tied together". For example in a simple 2-argument AF consisting of two arguments A and B mutually attacking each other, there may be two arguments but to all intents and purposes there is really only one "issue" at stake, and that is whether A or B (or neither) should be accepted. We want to isolate these different issues which are being argued over. Given an AF $\mathcal{A} = (Ar, \rightharpoonup)$, let us define the following two binary relations over Ar. For any $A, B \in Ar$:

- $A \equiv_1 B$ iff $\forall \mathcal{L} \in Comp_{\mathcal{A}} : \mathcal{L}(A) = \mathcal{L}(B)$
- $A \equiv_2 B$ iff $\forall \mathcal{L} \in Comp_{\mathcal{A}} : (\mathcal{L}(A) = \mathtt{in} \Leftrightarrow \mathcal{L}(B) = \mathtt{out}) \wedge (\mathcal{L}(A) = \mathtt{out} \Leftrightarrow \mathcal{L}(B) = \mathtt{in})$.

$A \equiv_1 B$ means that the labels assigned to A and B are exactly the same in all complete labellings, i.e., that A and B are in a sense logically equivalent, while $A \equiv_2 B$ means

that A and B always receive "opposite" labels: whenever A is labelled `in` then B is labelled `out`, and vice versa. It is easy to see that if $A \equiv_2 B$ then we also have $\mathcal{L}(A) = \texttt{undec}$ iff $\mathcal{L}(B) = \texttt{undec}$. From these two relations we define

$$A \equiv B \text{ iff } (A \equiv_1 B \vee A \equiv_2 B).$$

Thus if $A \equiv B$ then intuitively the labels of A and B are "in sync", in that the label of one cannot be changed without causing a change of equal magnitude to the label of the other.

PROPOSITION 6. *$\equiv$ is an equivalence relation over Ar.*

PROOF. *(Outline).* Reflexivity holds since $\equiv_1$ is reflexive. Symmetry holds since both $\equiv_1$ and $\equiv_2$ are symmetric, and transitivity holds because of the following composition properties: $(\equiv_1 \circ \equiv_1) = (\equiv_2 \circ \equiv_2) = \equiv_1$ and $(\equiv_1 \circ \equiv_2) = (\equiv_2 \circ \equiv_1) = \equiv_2$. □

Within each $\equiv$-equivalence class, there are at most 3 possible labellings which can occur: either *(i)* all its elements are labelled `undec`, or *(ii)* all its elements are set to `in` or `out`, or *(iii)* the "inverse" labelling to *(ii)* occurs, in which those arguments labelled `in` become `out` and those labelled `out` are now `in`. Essentially each equivalence class acts as a single 3-valued argument. We call each such class an *issue* of the given AF.

Definition 10. Given an AF $\mathcal{A} = (Ar, \rightharpoonup)$, the set $\mathbb{I}(\mathcal{A})$ of *issues* of $\mathcal{A}$ is defined as $\mathbb{I}(\mathcal{A}) = Ar/\equiv$. For $A \in Ar$ we will denote the $\equiv$-equivalence class of A by $[A]$.

For example, it can be checked that the issues for the AF in Fig. 3 are $\{A, B, C, D\}$, $\{E, F\}$ and $\{G, H\}$. In the AF of Fig. 2, however, there are 3 issues $\{A\}$, $\{B\}$ and $\{C\}$. Note that in the former case the issues coincide exactly with the *strongly connected components* [4] of the graph, whereas this is not true of the latter case.

Now, rather than calculate distance via argument-wise label difference as we did in Section 5, we can instead do it via *issue-wise* label difference. For this we need to define the measure of disagreement $DIFF(\mathcal{S}, \mathcal{T}, [A])$ between two labellings $\mathcal{S}$ and $\mathcal{T}$ on a single issue $[A]$. We do this using a basic label difference measure *diff*:

$$DIFF(\mathcal{S}, \mathcal{T}, [A]) \stackrel{\text{def}}{=} \mathit{diff}(\mathcal{S}(A), \mathcal{T}(A)).$$

PROPOSITION 7. *DIFF is well-defined, i.e., if $[A] = [B]$ then $\mathit{diff}(\mathcal{S}(A), \mathcal{T}(A)) = \mathit{diff}(\mathcal{S}(B), \mathcal{T}(B))$.*

PROOF. *(Outline)* If $[A] = [B]$ then either $A \equiv_1 B$ or $A \equiv_2 B$. In the former case the result is clear. In the latter case one may simply check for each of the 9 possibile combinations of labels for $\mathcal{S}(A)$, $\mathcal{T}(A)$. E.g., if $\mathcal{S}(A) = \texttt{in}$ and $\mathcal{T}(A) = \texttt{undec}$ then, from $A \equiv_2 B$ we know $\mathcal{S}(B) = \texttt{out}$ and $\mathcal{T}(B) = \texttt{undec}$. Thus $\mathit{diff}(\mathcal{S}(A), \mathcal{T}(A)) = \mathit{diff}(\mathcal{S}(B), \mathcal{T}(B))$ since *diff* by **(diff 4)**. □

Note that this result depends on the assumptions that *diff* satisfies **(diff 1)**, **(diff 2)** and **(diff 4)**.

Finally the issue-based distance measure *id* is defined by setting, for any $\mathcal{S}, \mathcal{T} \in \mathit{Comp}_{\mathcal{A}}$,

$$id(\mathcal{S}, \mathcal{T}) = \sum_{[A] \in \mathbb{I}(\mathcal{A})} DIFF(\mathcal{S}, \mathcal{T}, [A])$$

In the example in Fig. 3 we have $id(\mathcal{S}, \mathcal{T}_1) = 2 \times \mathit{diff}(\texttt{in}, \texttt{out})$ and $id(\mathcal{S}, \mathcal{T}_2) = \mathit{diff}(\texttt{in}, \texttt{out})$, as with the critical subsets approach of Section 6.1. For the example in Fig. 2 we get $id(\mathcal{L}_1, \mathcal{L}_2) = 2 \times \mathit{diff}(\texttt{in}, \texttt{out}) = id(\mathcal{L}_1, \mathcal{L}_3)$. So according to the issue-based distance $\mathcal{L}_2$ and $\mathcal{L}_3$ are equidistant from $\mathcal{L}_1$.

The issue-wise distance measure can be related to the preceding critical subsets approach. Clearly we have $id(\mathcal{S}, \mathcal{T}) = d_X(\mathcal{S}, \mathcal{T})$ (see equation (2) in Section 6.1), where X is any set formed by taking a representative of each $\equiv$-equivalence class. It turns out that we have the following:

PROPOSITION 8. *Let X be any set obtained by taking one element of each issue in $\mathbb{I}(\mathcal{A})$. Then $X \in \mathit{crit}(\mathcal{A})$.*

PROOF. *(Outline)* Let $\mathcal{L}_1$, $\mathcal{L}_2 \in \mathit{Comp}_{\mathcal{A}}$ be 2 complete labellings which agree on X. We must show $\mathcal{L}_1(A) = \mathcal{L}_2(A)$ for all $A \in Ar$. Let $A^* \in X$ be the chosen representative of $[A]$ in X, so $A^* \equiv A$. We know $\mathcal{L}_1(A^*) = \mathcal{L}_2(A^*)$. Denote this common label by `x`. If $A^* \equiv_1 A$ then both $\mathcal{L}_1(A)$ and $\mathcal{L}_2(A)$ are equal to `x` as required. Suppose $A^* \equiv_2 A$. If $\texttt{x} = \texttt{in}$ then both $\mathcal{L}_1(A)$ and $\mathcal{L}_2(A)$ are `out`. If $\texttt{x} = \texttt{out}$ then both $\mathcal{L}_1(A)$ and $\mathcal{L}_2(A)$ are `in`. Finally if $\texttt{x} = \texttt{undec}$ then both $\mathcal{L}_1(A)$ and $\mathcal{L}_2(A)$ are `undec`. □

Thus *id* can be thought of as a critical-set based distance which chooses from among a particular class of critical sets, viz. those which contain one argument from each issue. Furthermore, unlike the critical-set based distance the precise choice of these elements is irrelevant. However the critical set chosen need not be a minimal one, i.e., an element of $\mathit{mincrit}(\mathcal{A})$, as can be seen already in the AF of Fig. 2. We may deduce from all this and Theorem 1 the following:

THEOREM 2. *id satisfies* **(dm1)** *and* **(dm3)**. *Furthermore:*
(i). *If diff satisfies* **(diff 3+)** *then id satisfies* **(dm5)** *(and hence also* **(dm2)***).*
(ii). *If diff satisfies* **(diff 3+)** *and* **(diff 5)** *then id satisfies* **(dm5+)**.
(iii). *If diff satisfies* **(diff 3++)** *then id satisfies* **(dm4)**.

In addition, we have the following:

THEOREM 3. *The distance method $\mathcal{A} \mapsto id_{\mathcal{A}}$ satisfies* **(Iso)** *(and hence also* **(Auto)***).*

PROOF. *(Outline.)* Let $\mathcal{A}_1, \mathcal{A}_2$ be 2 AFs connected by isomorphism g and let $\mathcal{S}, \mathcal{T} \in \mathit{Comp}_{\mathcal{A}_1}$. To show $id_{\mathcal{A}_1}(\mathcal{S}, \mathcal{T}) = id_{\mathcal{A}_2}(g(\mathcal{S}), g(\mathcal{T}))$ we show that the summands on each side of this identity match up in pairs. More precisley we show there is a bijection $h : \mathbb{I}(\mathcal{A}_1) \to \mathbb{I}(\mathcal{A}_2)$ such that, for each $[A] \in \mathbb{I}(\mathcal{A}_1)$, $DIFF(\mathcal{S}, \mathcal{T}, [A]) = DIFF(g(\mathcal{S}), g(\mathcal{T}), h([A]))$. Indeed we can just define $h([A]) = [g(A)]$. The facts that h is well-defined and injective are both proved using the property that, for any $A, B \in Ar_1$, A and B belong to the same issue in $\mathbb{I}(\mathcal{A}_1)$ iff $g(A)$ and $g(B)$ belong to the same issue in $\mathbb{I}(\mathcal{A}_2)$. h is clearly surjective since, given any $[Z] \in \mathbb{I}(\mathcal{A}_2)$ we have $[Z] = h([g^{-1}(Z)])$. Finally $DIFF(g(\mathcal{S}), g(\mathcal{T}), h([A])) = DIFF(g(\mathcal{S}), g(\mathcal{T}), [g(A)]) = \mathit{diff}([g(\mathcal{S})](g(A)), [g(\mathcal{T})](g(A)))$. Since $[g(\mathcal{S})](g(A)) = \mathcal{S}(g^{-1}(g(A))) = \mathcal{S}(A)$ by definition of $g(\mathcal{S})$ (and similarly for $\mathcal{T}$), this equals $\mathit{diff}(\mathcal{S}(A), \mathcal{T}(A)) = DIFF(\mathcal{S}, \mathcal{T}, [A])$ as required. □

7. RELATED WORK AND CONCLUSION

We have initiated the investigation of the notion of distance between two reasonable evaluations of an argument graph. While this issue has been investigated in non-argument based accounts of both belief revision [18, 20], in judgement

aggregation [19, 21], and in abstract preferences [1], to our knowledge we are the first to study it in the context of formal argumentation theory.

We presented several different distance functions, all defined on top of a difference function on the space of possible labels $\{\texttt{in}, \texttt{out}, \texttt{undec}\}$. These functions fall into two groups: those which sum the difference between the labels of *all* arguments Ar in the framework, and those which single out various subsets of Ar as being in some sense the *critical* ones. We gave some postulates for such distance functions, even though we saw that many simple and straightforward candidates for distance measures suffer from some problem or another, and we developed some intuitions via several examples about what a distance function between complete labellings should be like.

For future work we would like to investigate more closely the issue-based distance method *id*. Specifically we are looking for other properties that it satisfies, perhaps leading to an *axiomatic characterisation*. We also want to apply these new distances to the problems of revision and judgement aggregation in argumentation. In revision we want to choose the closest labelling to the current one which extends an input new partial labelling. In questions of judgement aggregation we want to choose the labellings which are closest to the group as a whole. Similar considerations have been applied in propositional contexts (e.g. [13, 17]), while a first exploration of the use of Hamming-like distances (see Section 5) in labelling-aggregation has been carried out by Caminada et al. in [12], where it is used to check the manipulability and Pareto optimality of certain aggregation operators. They assume each member of a group of agents provides a complete labelling, and that each agent's preference relation over the set of all complete labellings is given by Hamming set or Hamming distance from its given labelling.

It would also be interesting to see if the issue-based methodology of Section 6.3 can be used to refine the distance-based approaches already existing in general judgement aggregation. Finally, here we focused on complete labellings. This is reasonable since they correspond to rational, coherent standpoints. But the definitions will work for other families of labellings too, like preferred, stable [14], and semi-stable [9].

8. ACKNOWLEDGEMENTS

Thanks are due to the reviewers for their encouraging remarks. Thanks also to Ringo Baumann, Gerhard Brewka and the Individual and Collective Reasoning group at the University of Luxembourg for some useful comments. Richard Booth is supported by the FNR/INTER project "Dynamics of Argumentation". Martin Caminada and Mikołaj Podlaszewski are supported by the National Research Fund, Luxembourg (FNR) (LAAMI and LAAMIcomp projects).

9. REFERENCES

[1] N. Baigent. Preference proximity and anonymous social choice. *The Quarterly Journal of Economics*, 102(1):161–169, 1987.

[2] P. Baroni, M. Caminada, and M. Giacomin. An introduction to argumentation semantics. *Knowledge Engineering Review*, 26:365–410, 2011.

[3] P. Baroni and M. Giacomin. On principle-based evaluation of extension-based argumentation semantics. *Artificial Intelligence*, 171(10-15):675–700, 2007.

[4] P. Baroni, M. Giacomin, and G. Guida. SCC-recursiveness: A general schema for argumentation semantics. *Artificial Intelligence*, 168(1-2):162–210, 2005.

[5] G. Brewka and J. Hertzberg. How to do things with worlds: On formalizing actions and plans. *Journal of Logic and Computation*, 3(5):517–532, 1993.

[6] M. Caminada. On the issue of reinstatement in argumentation. In *Proc. JELIA*, pages 111–123, 2006.

[7] M. Caminada. Comparing two unique extension semantics for formal argumentation: ideal and eager. In *Proc. BNAIC*, pages 81–87, 2007.

[8] M. Caminada and L. Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5-6):286–310, 2007.

[9] M. Caminada, W. Carnielli, and P. Dunne. Semi-stable semantics. *Journal of Logic and Computation*, 2011. in print.

[10] M. Caminada and D. Gabbay. A logical account of formal argumentation. *Studia Logica*, 93(2):109–145, 2009.

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

[12] M. Caminada, G. Pigozzi, and M. Podlaszewski. Manipulation in group argument evaluation. In *Proc. IJCAI*, pages 121–126, 2011.

[13] M. Dalal. Investigations into a theory of knowledge base revision. In *Proc. AAAI*, pages 475–479, 1988.

[14] 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):321–357, 1995.

[15] P. M. Dung, P. Mancarella, and F. Toni. Computing ideal sceptical argumentation. *Artificial Intelligence*, 171(10-15):642–674, 2007.

[16] D. Gabbay. Fibring argumentation frames. *Studia Logica*, 93(2):231–295, 2009.

[17] S. Konieczny, J. Lang, and P. Marquis. DA^2 merging operators. *Artificial Intelligence*, 157(1-2):49–79, 2004.

[18] D. Lehmann, M. Magidor, and K. Schlechta. Distance semantics for belief revision. *Journal of Symbolic Logic*, 66(1):295–317, 2001.

[19] M. Miller and D. Osherson. Methods for distance-based judgment aggregation. *Social Choice and Welfare*, 32(4):575–601, 2009.

[20] P. Peppas, S. Chopra, and N. Foo. Distance semantics for relevance-sensitive belief revision. In *Proc. KR*, pages 319–328, 2004.

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

[22] H. Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1(2):93–124, 2010.

[23] I. Rahwan and F. Tohmé. Collective argument evaluation as judgement aggregation. In *Proc. AAMAS*, pages 417–424, 2010.

[24] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.