On Diagnostic Arguments in Abstract Argumentation

Jordan Robinson[1,*], Katie Atkinson[1], Simon Maskell[1] and Chris Reed[2]

[1] *University of Liverpool, United Kingdom*
[2] *University of Dundee, Scotland*

Abstract

We are interested in employing argumentation for intelligence analysis due to its obvious potential for real-world impact. The Analysis of Competing Hypotheses, a well-known technique for multiple hypothesis evaluation from the intelligence community, includes sensitivity analysis as a task which helps analysts identify diagnostic information. We draw upon this notion of sensitivity analysis in this paper to set out a novel algorithm, called the *Diagnostic Argument Identifier*, that is able to identify *diagnostic* arguments. We employ a labelling-based approach to compute acceptance probabilities between partitions of arguments, which are then used to calculate the mutual information between the labels of each partition before and after the sequential removal of each argument from a framework. We present the results from one experiment on an abstract framework to assess whether our method can identify diagnostic arguments, and thus aid intelligence analysts. We argue that our algorithmic approach systematises and, therefore, reduces the subjectivity of sensitivity analysis; thus, yielding benefits to intelligence analysts – or any other expert working within a decision or deliberation setting – who need to objectively reevaluate the dependence of their set of conclusions on observed data present within an analysis.

Keywords

Abstract argumentation, probabilistic argumentation, information theory, intelligence analysis.

1. Introduction

Our work is conducted within the setting of intelligence analysis where one of the most well-known techniques within the intelligence community is the Analysis of Competing Hypotheses (ACH) [1]. The ACH provides a systematic approach to hypothesis evaluation. In short, the procedure starts with: hypothesis generation; listing of evidence; instantiation of a matrix with hypotheses (as column headers) and information, evidence and assumptions (as row headers); analysis of the consistency or otherwise of each row entry with each hypothesis; refinement of the matrix, removing or combining superfluous or overlapping hypotheses, respectively; drawing tentatively-held conclusions; sensitivity analysis to identify diagnostic row entries; and finally, reporting the probability of hypotheses and diagnostic row entries to stakeholders.

Sensitivity analysis forces intelligence analysts to think *diagnostically* in that they must establish whether the likelihood of their conclusions changes after the removal of a row entry. A *diagnostic* data point is one where its removal from the analysis changes the conclusions drawn. Although the ACH does its best to formalise hypothesis evaluation, sensitivity analysis is a challenging and subjective task because it is difficult to remove an observed data point and act as though it never existed when reevaluating.

SAFA'24: Fifth International Workshop on Systems and Algorithms for Formal Argumentation, September 17, 2024, Hagen, Germany

[*] Corresponding Author: Jordan Robinson; E-mail: J.Robinson9@liverpool.ac.uk.

To this end, we make two contributions. The first is an evaluation-based approach to probabilistic argumentation which uses the set of labellings discovered by a semantics to calculate joint and marginal argument acceptance probabilities of partitions of arguments and their labels. Second, we introduce the Diagnostic Argument Identifier (DAI), a novel algorithm, which applies the equations from the first contribution to quantify *diagnosticity scores* of arguments within an argumentation framework (AF), measuring the change in evaluation after the sequential removal of each argument from a framework. Our algorithm emulates the task of sensitivity analysis and should alleviate the reliance on human effort, through use of an algorithmic approach. In [2], we presented an application which executes the DAI and visualises the results output from the algorithm. The goal of this paper is to introduce the formal details underpinning the application.

The remainder of this paper is structured as follows. We begin by providing a brief overview of abstract argumentation in Section 2. In Section 3, we present our two main contributions: the joint and marginal argument acceptance probability equations and the DAI, which is accompanied by the relevant pseudo-code in Section 3.4. We present and discuss a result from a single experiment on an abstract example in Section 4. In Section 5, we relate the DAI to others' work, and Section 6 concludes with some avenues for future research.

2. Preliminaries

We consider (finite) Dung AFs $(\mathscr{A},\mathscr{R})$ containing a set of arguments $\mathscr{A}$ and attack relations $\mathscr{R} \subseteq \mathscr{A} \times \mathscr{A}$ [3]. For an AF $\mathscr{G}$ with arguments $\{a,b\} \subseteq \mathscr{A}$, we say that a attacks b if and only if $(a,b) \in \mathscr{R}$. The set of arguments that attack the argument $a \in \mathscr{A}$ is denoted by $Att(a) = \{b \mid (b,a) \in \mathscr{R}\}$. An admissible set of arguments has to satisfy *conflict-freeness* and *acceptability*. A set of arguments $S \subseteq \mathscr{A}$ is *conflict-free* if and only if $\nexists (a,b) \in \mathscr{R}$ where $a,b \in S$. An argument $a \in S \subseteq \mathscr{A}$ is *acceptable* with respect to S if and only if $\forall b \in \mathscr{A}$ such that $(b,a) \in \mathscr{R}$, then $\exists c \in S$ such that $(c,b) \in \mathscr{R}$.

We employ a labelling-based approach for this work because the set of labellings resulting from semantic evaluation enabled the computation of probabilities. A labelling $\mathscr{L}$ of a set of arguments $S \subseteq \mathscr{A}$ within an AF $\mathscr{G}$ is a total function $\mathscr{L}(S) : S \to \textsc{Lab}$ that assigns all the arguments $a \in S$ to a label $l \in \textsc{Lab}$, where $\textsc{Lab} = \{\textsc{in}, \textsc{out}, \textsc{und}\}$ in the case of complete semantics. We only consider complete labellings in this work as they are the foundation upon which all other semantics can be defined [4]. Let $\mathscr{G} = (\mathscr{A},\mathscr{R})$ be an AF and $\mathscr{L}(\mathscr{A}) : \mathscr{A} \to \textsc{Lab}$ be a labelling function. A *labelling* is a *complete labelling* if and only if $\forall a_i \in \mathscr{A}$, it holds that:

1. $\mathscr{L}(a_i) = \textsc{in}$ if and only if $\mathscr{L}(a_j) = \textsc{out}$, $\forall a_j$ such that $a_j \in Att(a_i)$;
2. $\mathscr{L}(a_i) = \textsc{out}$ if and only if $\exists a_j$ such that $a_j \in Att(a_i)$ and $\mathscr{L}(a_j) = \textsc{in}$.

As a consequence of a labelling being a total function, arguments that are neither labelled IN nor OUT are labelled UND.

Arguments labelled IN, OUT, or UND for one or all labellings are referred to as credulously or sceptically IN, OUT, or UND, respectively.

3. Diagnostic Argument Identifier

We now introduce our main contributions: the labelling-based argument acceptance probability equations for partitions of argument labels, derived from a set of probability spaces (Sec. 3.1), and then the DAI, which is an algorithm that is capable of identifying the most critical arguments within a Dung AF. We explain how to calculate the mutual information (MI) between partitions of labelling vectors (Sec. 3.2) and show how to conduct sensitivity analysis on an AF (Sec. 3.3). The section closes with our second contribution, the pseudo-code for the DAI (Sec. 3.4).

3.1. Probability Spaces

We let $\mathscr{G} = (\mathscr{A}, \mathscr{R})$ be an AF which contains a set of N arguments $\mathscr{A}$. We order the set of arguments using a function $f(\mathscr{A}) : \mathscr{A} \to \mathbf{A}$, where $\mathbf{A} = (a_1, ..., a_N)$ is an ordered vector of arguments. We assume we have a function $g(\mathbf{A}) : \mathbf{A} \to \mathscr{L}_M$ which assigns all arguments $a_i \in \mathbf{A}$ to a set of labelling vectors, such that

$$\mathscr{L}_M = \{\mathbf{L}_i\}_{i=1}^{M} \tag{1}$$

where M is the number of labelling vectors, and $\mathbf{L}_i$ is the i-th labelling vector containing N argument labels, such that $\mathbf{L}_i = (l_1, ..., l_N)$, where $l_j \in \mathbf{L}_i$ is the label of the argument $a_j \in \mathbf{A}$ and $l_j \in \textsc{Lab}$ (Sec. 2).

We now partition the set of arguments into two sets, named $\mathscr{A}_\phi$ and $\mathscr{A}_\psi$.

Definition 3.1. For an AF $\mathscr{G} = (\mathscr{A}, \mathscr{R})$, let the partitions of $\mathscr{A}$ be $\mathscr{A}_\phi \subseteq \mathscr{A}$ and $\mathscr{A}_\psi \subseteq \mathscr{A}$, where $\mathscr{A}_\phi \cup \mathscr{A}_\psi = \mathscr{A}$ and $\mathscr{A}_\phi \cap \mathscr{A}_\psi = \emptyset$, and the dichotomous sets $\mathscr{A}_\phi^c$ and $\mathscr{A}_\psi^c$ are complements such that $\mathscr{A}_\phi^c = \mathscr{A} \setminus \mathscr{A}_\psi$ and $\mathscr{A}_\psi^c = \mathscr{A} \setminus \mathscr{A}_\phi$.

Both the sets $\mathscr{A}_\phi$ and $\mathscr{A}_\psi$ are mapped to argument vectors through the function f, such that

$$f(\mathscr{A}_\phi) : \mathscr{A}_\phi \to \mathbf{A}_\phi = \{(a_1, ..., a_{|\mathscr{A}_\phi|}) \mid \forall a_i \in \mathscr{A} \text{ where } a_i \notin \mathscr{A}_\psi\} \text{ and} \tag{2}$$

$$f(\mathscr{A}_\psi) : \mathscr{A}_\psi \to \mathbf{A}_\psi = \{(a_1, ..., a_{|\mathscr{A}_\psi|}) \mid \forall a_j \in \mathscr{A} \text{ where } a_j \notin \mathscr{A}_\phi\}, \tag{3}$$

respectively, where $i \neq j$.

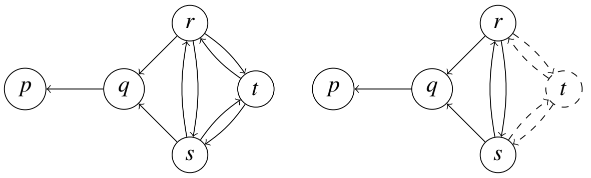
The partitions $\mathbf{A}_\phi$ and $\mathbf{A}_\psi$ are mapped to a corresponding set of labelling vectors through the same function g, such that $g(\mathbf{A}_\phi) : \mathbf{A}_\phi \to \mathscr{L}_\phi$ and $g(\mathbf{A}_\psi) : \mathbf{A}_\psi \to \mathscr{L}_\psi$, respectively, such that

$$\mathscr{L}_\phi = \{\mathbf{L}_{\phi,i}\}_{i=1}^{M_\phi} \text{ and} \tag{4}$$

$$\mathscr{L}_\psi = \{\mathbf{L}_{\psi,i}\}_{i=1}^{M_\psi} \tag{5}$$

where $M_\phi \leq M$, $M_\psi \leq M$, $\mathscr{L}_\phi \subseteq \mathscr{L}_M$, $\mathscr{L}_\psi \subseteq \mathscr{L}_M$, and $\mathbf{L}_{\phi,i}$ and $\mathbf{L}_{\psi,i}$ are the i-th labelling vectors for the partitions $\mathbf{A}_\phi$ and $\mathbf{A}_\psi$, respectively.

Example 3.1. Consider a Dung AF $\mathscr{G}$ with arguments $\mathscr{A} = \{p, q, r, s, t\}$ and relations $\mathscr{R} = \{(q, p), (r, q), (s, q), (r, s), (s, r), (r, t), (t, r), (s, t), (t, s)\}$ (Fig. 1a). Evaluating the AF $\mathscr{G}$ under complete semantics produces four distinct labellings (Tab. 1). Following Def. 3.1, we partition the set of arguments $\mathscr{A}$ into dichotomous sets, where $\mathscr{A}_\phi = \{p, q\}$ and $\mathscr{A}_\psi = \{r, s, t\}$. Using the



(a) $\mathscr{G}$ (b) $\mathscr{G}^t$

Figure 1: Example of a Dung AF

Table 1
The complete labellings of the AF displayed in Figure 1a.

Labellings, $\mathscr{L}(\mathscr{A})$	p	q	r	s	t
$\mathscr{L}_1(\mathscr{A})$	IN	OUT	IN	OUT	OUT
$\mathscr{L}_2(\mathscr{A})$	IN	OUT	OUT	IN	OUT
$\mathscr{L}_3(\mathscr{A})$	OUT	IN	OUT	OUT	IN
$\mathscr{L}_4(\mathscr{A})$	UND	UND	UND	UND	UND

function f, we map the sets of arguments $\mathscr{A}$, $\mathscr{A}_\phi$ and $\mathscr{A}_\psi$ to the argument vectors $\mathbf{A} = (p,q,r,s,t)$, $\mathbf{A}_\phi = (p,q)$ and $\mathbf{A}_\psi = (r,s,t)$, which enables the creation of sets of labelling vectors $\mathscr{L}_M$, $\mathscr{L}_\phi$ and $\mathscr{L}_\psi$, as shown in Eqs. 6, 7 and 8, respectively.

$$g(\mathbf{A}) : \mathbf{A} \rightarrow \mathscr{L}_M = \left\{ \begin{pmatrix} \text{IN} \\ \text{OUT} \\ \text{IN} \\ \text{OUT} \\ \text{OUT} \end{pmatrix}, \begin{pmatrix} \text{IN} \\ \text{OUT} \\ \text{OUT} \\ \text{IN} \\ \text{OUT} \end{pmatrix}, \begin{pmatrix} \text{OUT} \\ \text{IN} \\ \text{OUT} \\ \text{OUT} \\ \text{IN} \end{pmatrix}, \begin{pmatrix} \text{UND} \\ \text{UND} \\ \text{UND} \\ \text{UND} \\ \text{UND} \end{pmatrix} \right\} \tag{6}$$

$$g(\mathbf{A}_\phi) : \mathbf{A}_\phi \rightarrow \mathscr{L}_\phi = \left\{ \begin{pmatrix} \text{IN} \\ \text{OUT} \end{pmatrix}, \begin{pmatrix} \text{OUT} \\ \text{IN} \end{pmatrix}, \begin{pmatrix} \text{UND} \\ \text{UND} \end{pmatrix} \right\} \tag{7}$$

$$g(\mathbf{A}_\psi) : \mathbf{A}_\psi \rightarrow \mathscr{L}_\psi = \left\{ \begin{pmatrix} \text{IN} \\ \text{OUT} \\ \text{OUT} \end{pmatrix}, \begin{pmatrix} \text{OUT} \\ \text{IN} \\ \text{OUT} \end{pmatrix}, \begin{pmatrix} \text{OUT} \\ \text{OUT} \\ \text{IN} \end{pmatrix}, \begin{pmatrix} \text{UND} \\ \text{UND} \\ \text{UND} \end{pmatrix} \right\} \tag{8}$$

We let $(\Omega_\phi, \mathscr{F}_\phi, P_\phi)$ and $(\Omega_\psi, \mathscr{F}_\psi, P_\psi)$ be probability spaces, where Ω_ϕ and Ω_ψ are sample spaces, $\mathscr{F}_\phi$ and $\mathscr{F}_\psi$ are event spaces, and P_ϕ and P_ϕ are functions such that $P_\phi : \mathscr{F}_\phi \rightarrow (0,1]$ and $P_\psi : \mathscr{F}_\psi \rightarrow (0,1]$, respectively. We consider two random variables X_ϕ and X_ψ that are real-valued measurable functions $\mathrm{X}_\phi : \Omega_\phi \rightarrow \mathbb{R}$ and $\mathrm{X}_\psi : \Omega_\psi \rightarrow \mathbb{R}$ that map results from from the sample spaces Ω_ϕ and Ω_ψ to numerical values; thus, modelling a random experiment which, in our case, is the resulting set of labellings output from semantic evaluation of an AF.

Definition 3.2. (Random Vector). Let $(\Omega, \mathscr{F}, P)$ be a probability space where the random vector $\mathbf{X} : \Omega \rightarrow \mathbb{R}$ is a measurable function. The random vector $\mathbf{X}$ contains random variables $\mathbf{X} = (\mathrm{X}_\phi, \mathrm{X}_\psi)$ defined on two probability spaces $\mathrm{X}_\phi : \Omega_\phi \rightarrow \mathbb{R}$ and $\mathrm{X}_\psi : \Omega_\psi \rightarrow \mathbb{R}$.

Before semantic evaluation and given the constraints of argumentation, the number of possible elements (or labelling vectors) in the sample space Ω is the number of unique combinations of argument labels, such that

$$|\Omega| = |\textsc{Lab}|^N \tag{9}$$

where $|\Omega|$ is the number of potential elements in the sample space Ω, N is the number of arguments, and $\textsc{Lab}$ is defined in Sec. 2. After semantic evaluation of an AF, the set of labelling vectors in the sample space is reduced to a set of M labelling vectors, such that $\Omega = \mathscr{L}_M$.

The probability spaces $(\Omega_\phi, \mathscr{F}_\phi, P_\phi)$ and $(\Omega_\psi, \mathscr{F}_\psi, P_\psi)$ are measurable spaces, where $\mathscr{F}_\phi \otimes \mathscr{F}_\psi$ is the smallest σ-field of potential subsets of $\Omega_\phi \times \Omega_\psi \supseteq \Omega$, containing all sets of the form $\mathbf{L}_{\phi,i} \times \mathbf{L}_{\psi,j}$ where $\mathbf{L}_{\phi,i} \in \mathscr{F}_\phi$ and $\mathbf{L}_{\psi,j} \in \mathscr{F}_\psi$, such that $\mathscr{F}_\phi \otimes \mathscr{F}_\psi$ is the product σ-field. Thus, the event space $\mathscr{F}$ is a σ-algebra containing the powerset of all elements in the product space $\mathscr{F}_\phi \otimes \mathscr{F}_\psi$, including the empty set and the set of all events. However, after semantic evaluation, the set of combinations of labelling vectors for each partition is, again, reduced to sets of realised events such that $\mathscr{L}_\phi \otimes \mathscr{L}_\psi \subset \mathscr{F}_\phi \otimes \mathscr{F}_\psi$ is the smallest subset of events from which we can calculate non-zero probabilities, where $\mathscr{L}_\phi$ and $\mathscr{L}_\psi$ are shown in Eqs. 4 and 5, respectively.

The observation of the i-th outcome is denoted $x^{(i)} \in \Omega$ which is the i-th labelling vector $x^{(i)} \in \mathscr{L}_M$, such that $x^{(i)} = (x_\phi^{(i)}, x_\psi^{(i)}) = (\mathbf{L}_{\phi,i}, \mathbf{L}_{\psi,i}) \in \Omega$. Similarly, we can observe outcomes from different labellings, where $x_\phi^{(i)}, x_\psi^{(j)} \in \Omega$ are the i-th and j-th labelling vector for the partitions $\mathbf{A}_\phi$ and $\mathbf{A}_\psi$, respectively, where $i \neq j$. With a slight abuse of notation, the probability of $\mathbf{A}_\phi$ and $\mathbf{A}_\psi$'s labelling vectors being contained in the i-th and j-th outcome in the sample space is denoted as $P(x_\phi^{(i)})$ or $P(x_\psi^{(j)})$ which refers to $P(\mathrm{X}_\phi = x_\phi^{(i)})$ or $P(\mathrm{X}_\psi = x_\psi^{(j)})$, respectively.

The joint and marginal probabilities of argument labels in each segment are the only probabilities concern our calculation of the MI for each combination of $\mathbf{A}_\phi$ and $\mathbf{A}_\psi$.

3.1.1. Joint Probability

First, the joint probability of the labels of arguments in each partition measured across the i-th and j-th outcome in the sample space Ω is computed using Eq. 10.

$$P(x_\phi^{(i)},\, x_\psi^{(j)}) = \frac{1}{M} \sum_{k=1}^{M} \mathbb{I}_{x_\phi^{(i)} = x_\phi^{(k)};\, x_\psi^{(j)} = x_\psi^{(k)}} \tag{10}$$

where $\mathbb{I}_A$ is unity if A is true and zero otherwise.

Due to the distinct nature of labellings, there will only be one labelling vector in the sample space Ω that contains the same arrangement of argument labels as the labels for arguments in each partition. Thus, there will be M pairs of labellings vectors which produce non-zero joint probabilities across the product space of realised argument labels in each segment. Consider a list $\mathscr{M}$ containing $|\mathscr{M}|$ distinct i-j pairs, where $|\mathscr{M}| = M_\phi \times M_\psi$, from the observed product space $\mathscr{L}_\phi \otimes \mathscr{L}_\psi$, such that we compute $|\mathscr{M}|$ joint probabilities. Each joint probability across the product space is the reciprocal of the number of labellings, if and only if both observed outcomes feature in the same labelling vector across the sample space Ω, and for any other i-j pair the probability is zero, as shown in Eq. 11.

$$P(x_\phi^{(i)},\, x_\psi^{(j)}) = \begin{cases} \frac{1}{M}, & \text{iff } \exists m \in \{1, \ldots, M\} \text{ such that } i = i(m) \text{ and } j = j(m) \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

It follows that the list $\mathscr{M}$ contains M i-j pairs of labelling vectors with a joint probability greater than zero, corresponding to the number of times that the sample space Ω contained those distinct segments of argument labels, or outcomes $x^{(m)} = (x_\phi^{(m)}, x_\psi^{(m)}) \in \Omega$.

Example 3.2. Continuing our running example, we employ the sets of labelling vectors $\mathscr{L}_\phi$ and $\mathscr{L}_\psi$, presented in Eqs. 7 and 8, to compute the joint probability of labels for arguments in each partition. The product space $\mathscr{L}_\phi \otimes \mathscr{L}_\psi$ contains $|\mathscr{M}| = M_\phi \times M_\psi$ combinations of potential labelling vectors based on the unique vectors in each partition, where $|\mathscr{M}| = 12$ in this example. Using both Eqs. 10 and 11, it is easy to see that the first, second and last i-j pair from the product space $\mathscr{L}_\phi \otimes \mathscr{L}_\psi$, shown in Eq. 12, is equal to $\frac{1}{4}$, which is $\frac{1}{M}$ in this example. The joint probability of the third i-j pair in Eq. 12 is equal to zero in both Eqs. 10 and 11 because that event was not observed as a labelling vector in the sample space Ω (see Tab. 1).

$$\mathscr{L}_\phi \otimes \mathscr{L}_\psi = \left\{ \begin{pmatrix} \text{IN} \\ \text{OUT} \\ \text{IN} \\ \text{OUT} \\ \text{OUT} \end{pmatrix}, \begin{pmatrix} \text{IN} \\ \text{OUT} \\ \text{OUT} \\ \text{IN} \\ \text{OUT} \end{pmatrix}, \begin{pmatrix} \text{IN} \\ \text{OUT} \\ \text{OUT} \\ \text{OUT} \\ \text{IN} \end{pmatrix}, \ldots, \begin{pmatrix} \text{UND} \\ \text{UND} \\ \text{UND} \\ \text{UND} \\ \text{UND} \end{pmatrix} \right\} \tag{12}$$

3.1.2. Marginal Probability

The second probability that we wish to measure is the marginal probability of each unique labelling vector in the sets $\mathscr{L}_\phi$ and $\mathscr{L}_\psi$, which turns out to be a normalised count across a vector subspace, counting how many times $\mathrm{X}_\phi = x_\phi^{(i)}$ and $\mathrm{X}_\psi = x_\psi^{(i)}$ occurred in the sample space Ω, as shown in Eqs. 13 and 14, respectively.

$$P(x_\phi^{(i)}) = \frac{1}{M} \sum_{j=1}^{M} \mathbb{I}_{x_\phi^{(i)} = x_\phi^{(j)}} \tag{13}$$

$$P(x_\psi^{(i)}) = \frac{1}{M} \sum_{j=1}^{M} \mathbb{I}_{x_\psi^{(i)} = x_\psi^{(j)}} \tag{14}$$

Example 3.3. Consider again the running example from Fig. 1a and the partition $\mathbf{A}_\phi$ with its set of distinct labelling vectors $\mathscr{L}_\phi$, as presented in Eq. 7. Using Eq. 13, there are three events within the space $\mathscr{L}_\phi$ which we can calculate marginal probabilities for. The marginal probability of the first $\mathbf{L}_{\phi,1} \in \Omega$, second $\mathbf{L}_{\phi,2} \in \Omega$ and third event $\mathbf{L}_{\phi,3} \in \Omega$ are equal to $\frac{1}{2}$, $\frac{1}{4}$ and $\frac{1}{4}$, respectively.

3.2. Mutual Information

We wish to quantify the amount of information in the initial AF and the extent to which removing an argument affects the distribution of labels between partitions of other arguments. We measure this change through the computation of the MI between partitions of argument labels before and after the removal of an argument. The MI is a symmetric function that quantifies the conditional dependence between two random variables and is able to determine the amount of information communicated, on average, about one random variable through observation of another [5, 6]. The

MI between pairs of argument labels across the realised product space $\mathscr{L}_\phi \otimes \mathscr{L}_\psi$, where $\mathscr{L}_\phi$ and $\mathscr{L}_\psi$ are defined Eqs. 4 and 5, respectively, is shown in Eq. 15.

$$I(\mathrm{X}_\phi;\mathrm{X}_\psi) = \sum_{x_\phi^{(i)} \in \mathscr{L}_\phi} \sum_{x_\psi^{(j)} \in \mathscr{L}_\psi} P(x_\phi^{(i)}, x_\psi^{(j)}) \log \left(\frac{P(x_\phi^{(i)}, x_\psi^{(j)})}{P(x_\phi^{(i)})P(x_\psi^{(j)})} \right) \tag{15}$$

The units of MI depend on the base of the logarithm used in the calculation. For the purposes of this paper, information will be measured in *bits* which is the logarithm to the base of two.

We simplify the MI calculation presented in Eq. 15 by conducting one summation over $|\mathscr{M}|$ pairs of outcomes $x_\phi^{(i)}$ and $x_\psi^{(j)}$ which feature in the realised product space $\mathscr{L}_\phi \otimes \mathscr{L}_\psi$, where $|\mathscr{M}| = M_\phi \times M_\psi$, as shown in Eq. 16.

$$I(\mathrm{X}_\phi;\mathrm{X}_\psi) = \sum_{(x_\phi^{(i)}, x_\psi^{(j)}) \in \mathscr{L}_\phi \otimes \mathscr{L}_\psi} P(x_\phi^{(i)}, x_\psi^{(j)}) \log \left(\frac{P(x_\phi^{(i)}, x_\psi^{(j)})}{P(x_\phi^{(i)})P(x_\psi^{(j)})} \right) \tag{16}$$

We note that many of the joint probabilities in this sum will be zero for *i*-*j* pairs of outcomes in the product space that do not feature in the sample space. So, the only non-zero contributions to the MI summation will be *i*-*j* pairs that are in both the observed sample space Ω and the product space $\mathscr{L}_\phi \otimes \mathscr{L}_\psi$. Thus, we restrict the summation to only include outcomes $x_\phi^{(i)}$ and $x_\psi^{(j)}$ that have a joint probability greater than zero, as in Eq. 17.

$$I(\mathrm{X}_\phi;\mathrm{X}_\psi) = \sum_{(x_\phi^{(i)}, x_\psi^{(j)}) \in \mathscr{L}_\phi \otimes \mathscr{L}_\psi, P(x_\phi^{(i)}, x_\psi^{(j)})>0} P(x_\phi^{(i)}, x_\psi^{(j)}) \log \left(\frac{P(x_\phi^{(i)}, x_\psi^{(j)})}{P(x_\phi^{(i)})P(x_\psi^{(j)})} \right) \tag{17}$$

In light of this constraint and using Eq. 11, it is easy to see that the only non-zero addends to the MI will be from contributions where the joint probability of pairs of outcomes from the product space feature in the same labelling from the sample space. For the sake of computational efficiency, we substitute Eq. 11 into Eq. 17 so that the MI calculation is, therefore, reduced to a single sum over the set of labellings for our problem setting, such that

$$I(\mathrm{X}_\phi;\mathrm{X}_\psi) = \sum_{x^{(m)} \in \Omega}^{M} \frac{1}{M} \log \left(\frac{1}{M\, P(x_\phi^{(m)})P(x_\psi^{(m)})} \right) \tag{18}$$

where $x^{(m)} = (x_\phi^{(m)}, x_\psi^{(m)})$ and $\Omega \subset \mathscr{L}_\phi \otimes \mathscr{L}_\psi$ is the space containing non-zero joint probabilities.

The MI is a symmetric function which means there will be a total of $\frac{2^{\mathscr{A}}}{2}$ distinct combinations of $\mathbf{A}_\phi$ and $\mathbf{A}_\psi$ which produce unique MI calculations. We, therefore, compute the MI between divisions of arguments for up to half the powerset to completely explore the distribution of information communicated across the sets of labelling vectors within the initial AF.

Example 3.4. Turning back to our running example and remembering that $\mathbf{A}_\phi = (p,q)$ and $\mathbf{A}_\psi = (r,s,t)$. The MI between each segment's set of labelling vectors was found to be $I(\mathrm{X}_\phi;\mathrm{X}_\psi) = 1.5$ *bits* (1 d.p.), under complete semantics. We see that observing the labels of, say, the arguments in $\mathbf{A}_\phi$ tells us 1.5 *bits* of information about the labels of arguments in $\mathbf{A}_\psi$, and vice versa.

3.3. Sensitivity Analysis

Now that we have explained how to determine the MI between segments of labelling vectors within the initial AF, we explain how to conduct sensitivity analysis. To start this task, we sequentially remove each argument $a \in \mathscr{A}$ and the relations containing that argument from the initial AF – creating a sensitive AF $\mathscr{G}^a$ – which we evaluate using the same semantics chosen earlier, and compute the MI between divisions of labelling vectors in the sensitive AF.

Definition 3.3. For an AF $\mathscr{G} = (\mathscr{A}, \mathscr{R})$ undergoing sensitivity analysis, we refer to $\mathscr{G}^a = (\mathscr{A}^a, \mathscr{R}^a)$ as a sensitive AF which does not include the argument a[1], the argument of interest, where $\mathscr{A}^a = \mathscr{A} \setminus a$ such that $\mathscr{A}^a \subset \mathscr{A}$ and $a \notin \mathscr{A}^a$. For a relation $r \in \mathscr{R}$, if the argument $a \in r$ features in that relation, then it is removed from the set of relations such that $\mathscr{R}^a = \mathscr{R} \setminus \mathscr{R}^-$ is the set of sensitive relations and $\mathscr{R}^- = \{\, r \mid \forall r \in \mathscr{R} \text{ where } a \in r\}$ are the relations to be removed, where $\mathscr{R}^a \subseteq \mathscr{R}$.

Notation 3.1. When we say $\mathscr{A}^a_\phi$ or $\mathscr{A}^a_\psi$, we are referring to the two partitions of $\mathscr{A}^a$ within the sensitive AF $\mathscr{G}^a$ that obey Def. 3.1, however they use the set of sensitive arguments $\mathscr{A}^a$, as outlined in Def. 3.3, instead of the set of all arguments $\mathscr{A}$ in the initial AF.

Again, we order the set of sensitive arguments $\mathscr{A}^a$ using a function $f(\mathscr{A}^a) : \mathscr{A}^a \to \mathbf{A}^a$, such that $\mathbf{A}^a = (a_1, ..., a_{N-1})$ is a vector of arguments. We map the vector of sensitive arguments $\mathbf{A}^a$ to a set of labelling vectors through the function $g(\mathbf{A}^a) : \mathbf{A}^a \to \mathscr{L}^a_M$, such that

$$\mathscr{L}^a_M = \{\mathbf{L}^a_i\}_{i=1}^{M^a} \tag{19}$$

where M^a is the number of labelling vectors output from semantic evaluation of the sensitive AF $\mathscr{G}^a$, and $\mathbf{L}^a_i$ is the i-th labelling vector containing $N-1$ argument labels, such that $\mathbf{L}^a_i = (l^a_1, ..., l^a_{N-1})$, where $l^a_j \in \mathbf{L}^a_i$ is the label of the argument $a_j \in \mathbf{A}^a$ and $l^a_j \in \textsc{Lab}$ (Sec. 2).

The sets $\mathscr{A}^a_\phi$ and $\mathscr{A}^a_\psi$ are mapped to the argument vectors $\mathbf{A}^a_\phi$ and $\mathbf{A}^a_\phi$ through

$$f(\mathscr{A}^a_\phi) : \mathscr{A}^a_\phi \to \mathbf{A}^a_\phi = \{(a_1, ..., a_{|\mathscr{A}^a_\phi|}) \mid \forall a_i \in \mathscr{A}^a \text{ where } a_i \notin \mathscr{A}^a_\psi\} \text{ and} \tag{20}$$

$$f(\mathscr{A}^a_\psi) : \mathscr{A}^a_\psi \to \mathbf{A}^a_\psi = \{(a_1, ..., a_{|\mathscr{A}^a_\psi|}) \mid \forall a_j \in \mathscr{A}^a \text{ where } a_j \notin \mathscr{A}^a_\phi\}, \tag{21}$$

respectively, where $i \neq j$.

The sets of distinct labelling vectors corresponding to the partitions $\mathbf{A}_\phi$ and $\mathbf{A}_\psi$ are found using

$$g(\mathbf{A}^a_\phi) : \mathbf{A}^a_\phi \to \mathscr{L}^a_\phi = \{\mathbf{L}^a_{\phi,i}\}_{i=1}^{M^a_\phi} \text{ and} \tag{22}$$

$$g(\mathbf{A}^a_\psi) : \mathbf{A}^a_\psi \to \mathscr{L}^a_\psi = \{\mathbf{L}^a_{\psi,i}\}_{i=1}^{M^a_\psi} \tag{23}$$

where $M^a_\phi \leq M^a$, $M^a_\psi \leq M^a$, $\mathscr{L}^a_\phi \subseteq \mathscr{L}^a_M$, $\mathscr{L}^a_\psi \subseteq \mathscr{L}^a_M$, and $\mathbf{L}^a_{\phi,i}$ and $\mathbf{L}^a_{\psi,i}$ are the i-th labelling vectors for the partitions $\mathbf{A}^a_\phi$ and $\mathbf{A}^a_\psi$, respectively.

Using the same semantics as earlier, we evaluate the sensitive AF to observe two sets of labelling vectors for the arguments in the vectors $\mathbf{A}^a_\phi$ and $\mathbf{A}^a_\psi$; thus, enabling us to compute argument acceptance probabilities for partitions of arguments within the sensitive AF.

[1] It is important for the reader to note that this paper only considers the removal of one argument from an initial AF while conducting sensitivity analysis. However, we note this approach could be extended to remove more than one argument from an initial AF to understand how this affects the results output from sensitivity analysis.

Notation 3.2. Let $\mathscr{G}^a$ be a sensitive AF with the argument a removed (Def. 3.3) and $\mathbf{A}^a_\phi$ and $\mathbf{A}^a_\psi$ be two sensitive argument vectors, mapped from the sets $\mathscr{A}^a_\phi$ and $\mathscr{A}^a_\psi$, which follow Def. 3.1. The random vector $\tilde{\mathbf{X}} = (\tilde{\mathrm{X}}_\phi, \tilde{\mathrm{X}}_\psi)$ is a measurable function from a probability space $(\Omega^a, \mathscr{F}^a, P^a)$ where $\tilde{\mathbf{X}}$ maps elements from the sensitive sample space $\Omega^a = \mathscr{L}^a_M$ to events $\mathscr{L}^a_\phi \otimes \mathscr{L}^a_\psi \subset \mathscr{F}^a$.

Example 3.5. Back to our running example. We present an MI calculation for the sensitive AF $\mathscr{G}^t$. To start, we employ Def. 3.3 to create a sensitive AF $\mathscr{G}^t$ with arguments $\mathscr{A}^t = \{p,q,r,s\}$ and relations $\mathscr{R}^t = \{(q,p),(r,q),(s,q),(r,s),(s,r)\}$ (Fig. 1b). We let $\mathscr{A}^t_\phi = \{p,q\}$ and $\mathscr{A}^t_\psi = \{r,s\}$ because that is the same as the original partitions (i.e., $\mathscr{A}_\phi$ and $\mathscr{A}_\psi$ in Ex. 3.1) where neither $\mathscr{A}^t_\psi$ nor $\mathscr{A}^t_\psi$ contain the argument t. We map the dichotomous sets of arguments $\mathscr{A}^t_\phi$ and $\mathscr{A}^t_\psi$ to the argument vectors $\mathbf{A}^t_\phi = (p,q)$ and $\mathbf{A}^t_\psi = (r,s)$, respectively, through the function f. We, again, use complete semantics to evaluate $\mathscr{G}^t$ to observe the set labellings for each partition (Tab. 2), allowing the realisation and mapping of each argument in $\mathbf{A}^t_\phi$ and $\mathbf{A}^t_\psi$ to their respective sets of labelling vectors $\mathscr{L}^t_\phi$ and $\mathscr{L}^t_\psi$, through the function g. The MI $I(\tilde{\mathrm{X}}_\phi;\tilde{\mathrm{X}}_\psi)$ between the labelling vectors for this partition of sensitive arguments is equal to 0.9 *bits* (1 d.p.).

Table 2
The complete labellings of the $\mathscr{G}^t$ displayed in Figure 1b.

Labellings, $\mathscr{L}(\mathscr{A}^t)$	p	q	r	s
$\mathscr{L}_1(\mathscr{A}^t)$	IN	OUT	IN	OUT
$\mathscr{L}_2(\mathscr{A}^t)$	IN	OUT	OUT	IN
$\mathscr{L}_3(\mathscr{A}^t)$	UND	UND	UND	UND

For a sensitive AF $\mathscr{G}^a$, the total number of combinations of $\mathscr{A}^a_\phi$ and $\mathscr{A}^a_\psi$ that produce unique MI calculations is equal to $\frac{2^{\mathscr{A}^a}}{2}$, which is equal to half the number of unique MI results (i.e., $\frac{2^{\mathscr{A}}}{2}$) that we calculate in the initial AF $\mathscr{G}$.

Once we have computed the set of distinct MI results using half the possible combinations of $\mathscr{A}_\phi$ and $\mathscr{A}_\psi$ in the initial AF and $\mathscr{A}^a_\phi$ and $\mathscr{A}^a_\psi$ in the sensitive one, we can quantify a *diagnosticity score*, which describes how much information was lost or gained between either $\mathscr{A}_\phi$ and $\mathscr{A}_\psi$'s set of labelling vectors after the removal of the sensitive argument, as stated in Def. 3.4.

Definition 3.4. The *diagnosticity score* is defined as the change in MI before (calculated using the labelling vectors $\mathscr{L}_\phi$ and $\mathscr{L}_\psi$ for all arguments in $\mathscr{A}_\phi$ and $\mathscr{A}_\psi$ from $\mathscr{G}$) and after (calculated using $\mathscr{L}^a_\phi$ and $\mathscr{L}^a_\psi$ for all arguments in $\mathscr{A}^a_\phi$ and $\mathscr{A}^a_\psi$ from $\mathscr{G}^a$) the removal of the argument of interest. Eq. 24 computes the *diagnosticity score* when the argument of interest a was removed from either $\mathscr{A}_\phi$ or $\mathscr{A}_\psi$.

$$\mathscr{D}(\mathscr{A}_\phi, \mathscr{A}_\psi; \mathscr{A}^a_\phi \vee \mathscr{A}^a_\psi) = I(\mathrm{X}_\phi;\mathrm{X}_\psi) - I(\tilde{\mathrm{X}}_\phi;\tilde{\mathrm{X}}_\psi) \tag{24}$$

An interesting point to note about Eq. 24 is the sign. A positive diagnosticity score infers that there was more information communicated, on average, between partitions in the initial AF $\mathscr{G}$, whereas a score below zero indicates that there was more information transferred between partitions of argument labels within the sensitive AF $\mathscr{G}^a$.

Example 3.6. Coming back to our running example for the last time, we consider the impact that removing the argument t had on the acceptability of arguments within each partition in the initial and sensitive AFs. We employed the set of labelling vectors of $\mathbf{A}_\phi$ and $\mathbf{A}_\psi$ to calculate the MI between X_ϕ and X_ψ in the initial AF (Ex. 3.4), and used the set of labelling vectors of $\mathbf{A}^t_\phi$ and $\mathbf{A}^t_\psi$ to compute the MI between $\tilde{\mathrm{X}}_\phi$ and $\tilde{\mathrm{X}}_\psi$ in the sensitive AF $\mathscr{G}^t$ (Ex. 3.5). We now show the use Def. 3.4 by presenting the diagnosticity score $\mathscr{D}(\mathscr{A}_\phi, \mathscr{A}_\psi; \mathscr{A}^t_\psi)$ for the aforementioned partitions, which turns out to be 0.6 *bits* (1 d.p.).

3.4. Pseudo-code for the Diagnostic Argument Identifier

We present the pseudo-code for the DAI in Algorithm 1. The algorithm takes as input an AF $\mathscr{G}$ and a semantics $\mathcal{S}$ capable of producing more than one labelling, and returns a *diagnosticity vector* $\mathrm{D}[\mathscr{A}][\frac{2^{\mathscr{A}}}{2}]$, which contains $\frac{2^{\mathscr{A}}}{2}$ diagnosticity scores for every argument in the graph.

4. Experimental Evaluation

We have introduced a novel algorithm that uses an evaluation-based approach to quantify how much the removal of an argument changes the semantic evaluation of an AF (Algo. 1). To evaluate the effectiveness of our method, we now present the results of an experiment conducted using the AF in our running example (Fig. 1). One of the reasons we chose to use the AF presented in Fig. 1 was because we assumed it possessed a similar topology to graphs found within an intelligence setting (i.e., many symmetric attacks as a result of conflicting information).

Results and Discussion. We computed $\frac{2^{\mathscr{A}}}{2}$ (i.e., 16) diagnosticity scores, corresponding to the total number of segments of labelling vectors, for all arguments in the running example (Fig. 2).

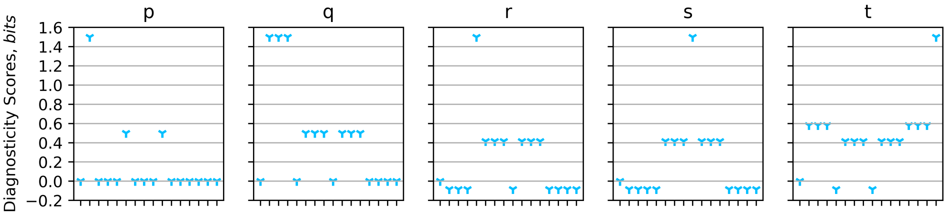


Figure 2: The diagnosticity scores for the 16 partitions of labelling vectors for all arguments within the AF $\mathscr{G}$ (Fig. 1), discovered using complete semantics.

Looking to Fig. 3, we present a violin plot which groups together the individual diagnosticity scores in Fig. 2 to show the distribution of change in MI for partitions of labelling vectors after the sequential removal of each argument in the AF $\mathscr{G}$ within the running example. We also included the median and mean diagnosticity scores, as well as the average absolute values, to better understand and compare the distribution and sign (i.e., positive or negative) of scores for all the removed arguments.

Algorithm 1 The Diagnostic Argument Identifier

```
Input: An AF $\mathscr{G} = (\mathscr{A}, \mathscr{R})$ and a semantics $\mathcal{S}$
Output: $\mathrm{D}[\mathscr{A}][\frac{2^{\mathscr{A}}}{2}]$
function DAI($\mathscr{G}$, $\mathcal{S}$)
    $\mathscr{L}(\mathscr{A})$    ▷ Evaluate initial AF $\mathscr{G}$ with chosen semantics $\mathcal{S}$.
    $\mathbf{A} = f(\mathscr{A}) : \mathscr{A} \to \mathbf{A}$    ▷ Create a vector of arguments for all arguments in the initial AF.
    $\Omega = g(\mathbf{A}) : \mathbf{A} \to \mathscr{L}_M$    ▷ Set the sample space to be equal to the set of labelling vectors from the initial AF.
    $M = |\mathscr{L}_M|$    ▷ Save the number of labellings from the initial AF.
    $\mathrm{MIs}[\frac{2^{\mathscr{A}}}{2}]$    ▷ Declare the size of the initial AF's MI results array.
    $\mathrm{D}[\mathscr{A}][\frac{2^{\mathscr{A}}}{2}]$    ▷ Declare the size of *diagnosticity vector D*.
    for $\mathscr{A}_\phi \in \frac{2^{\mathscr{A}}}{2}$ do    ▷ Iterate through subsets of $\mathscr{A}_\phi$ in the initial AF, up to half the powerset (Def. 3.1).
        $\mathscr{A}_\psi = \mathscr{A} \setminus \mathscr{A}_\phi$    ▷ Create the other set $\mathscr{A}_\psi$ (Def. 3.1).
        $\mathbf{A}_\phi = f(\mathscr{A}_\phi) : \mathscr{A}_\phi \to \mathbf{A}_\phi$;   $\mathbf{A}_\psi = f(\mathscr{A}_\psi) : \mathscr{A}_\psi \to \mathbf{A}_\psi$    ▷ Create the argument vectors for $\mathscr{A}_\phi$ and $\mathscr{A}_\psi$.
        $\mathscr{L}_\phi = g(\mathbf{A}_\phi) : \mathbf{A}_\phi \to \mathscr{L}_\phi$;   $\mathscr{L}_\psi = g(\mathbf{A}_\psi) : \mathbf{A}_\psi \to \mathscr{L}_\psi$ ▷ Create the sets of labelling vectors (Eqs. 4 & 5).
        $I(\mathrm{X}_\phi;\mathrm{X}_\psi) = 0$    ▷ Set the MI equal to zero for the current partitions of the initial AF.
        for $(x_\phi^{(m)}, x_\psi^{(m)}) \in \Omega$ do    ▷ Iterate through labellings in the $\Omega$ where both $x_\phi^{(m)}$ and $x_\psi^{(m)}$ occur.
            $P(x_\phi^{(m)})$; $P(x_\psi^{(m)})$    ▷ Eqs. 13 and 14, respectively.
            $I(\mathrm{X}_\phi;\mathrm{X}_\psi) \mathrel{+}= \frac{1}{M}\log\left(\frac{1}{M\,P(x_\phi^{(m)})P(x_\psi^{(m)})}\right)$    ▷ Compute the addend of the MI between segments for the $m$-th labelling vector of the initial AF (Eq. 18).
        $\mathrm{MIs} \leftarrow \{\mathscr{A}_\phi,\ \mathscr{A}_\psi,\ I(\mathrm{X}_\phi;\mathrm{X}_\psi)\}$    ▷ Append the result to the array of initial MI calculations.
    for $a \in \mathscr{A}$ do    ▷ Begin sensitivity analysis where $a$ is the *argument of interest*.
        $\mathscr{G}^a = (\mathscr{A}^a, \mathscr{R}^a)$    ▷ Instantiate the sensitive AF (Def. 3.3).
        $\mathscr{L}(\mathscr{A}^a)$    ▷ Evaluate the sensitive AF $\mathscr{G}^a$ with chosen semantics $\mathcal{S}$.
        $\mathbf{A}^a = f(\mathscr{A}^a) : \mathscr{A}^a \to \mathbf{A}^a$    ▷ Create the argument vector for the set of all sensitive arguments.
        $\Omega^a = g(\mathbf{A}) : \mathbf{A} \to \mathscr{L}_M^a$    ▷ Set the sample space to be equal to the sensitive AF's set of labelling vectors.
        $M^a = |\mathscr{L}_M^a|$    ▷ Save the number of labellings of the sensitive AF.
        for $\mathscr{A}_\phi$, $\mathscr{A}_\psi$, $I(\mathscr{A}_\phi;\mathscr{A}_\psi,) \in MIs$ do    ▷ Iterate through partitions and MI results from the initial AF $\mathscr{G}$.
            if $a \in \mathscr{A}_\phi$ then
                $\mathscr{A}_\phi^a = \mathscr{A}_\phi \setminus a$; $\mathscr{A}_\psi^a = \mathscr{A}_\psi$    ▷ Create $\mathscr{A}_\phi^a$ and $\mathscr{A}_\psi^a$ by removing $a$ from $\mathscr{A}_\phi$ (Defs. 3.1 & 3.3).
            else if $a \in \mathscr{A}_\psi$ then
                $\mathscr{A}_\psi^a = \mathscr{A}_\psi \setminus a$; $\mathscr{A}_\phi^a = \mathscr{A}_\phi$    ▷ Create $\mathscr{A}_\phi^a$ and $\mathscr{A}_\psi^a$ by removing $a$ from $\mathscr{A}_\psi$ (Defs. 3.1 & 3.3).
            $\mathbf{A}_\phi^a = f(\mathscr{A}_\phi^a) : \mathscr{A}_\phi^a \to \mathbf{A}_\phi^a$; $\mathbf{A}_\psi^a = f(\mathscr{A}_\psi^a) : \mathscr{A}_\psi^a \to \mathbf{A}_\psi^a$    ▷ Create the sensitive AF's argument vectors.
            $\mathscr{L}_\phi^a = g(\mathbf{A}_\phi^a) : \mathbf{A}_\phi^a \to \mathscr{L}_\phi^a$; $\mathscr{L}_\psi^a = g(\mathbf{A}_\psi^a) : \mathbf{A}_\psi^a \to \mathscr{L}_\psi^a$    ▷ Create the sensitive sets of labelling vectors.
            $I(\tilde{\mathrm{X}}_\phi;\tilde{\mathrm{X}}_\psi) = 0$    ▷ Set the MI equal to zero for the current partitions of the sensitive AF.
            for $(\tilde{x}_\phi^{(\tilde{m})}, \tilde{x}_\psi^{(\tilde{m})}) \in \Omega^a$ do    ▷ Iterate through the sensitive sample space $\Omega^a$
                $P(\tilde{x}_\phi^{(\tilde{m})})$; $P(\tilde{x}_\psi^{(\tilde{m})})$    ▷ Eqs. 13 and 14, respectively.
                $I(\tilde{\mathrm{X}}_\phi;\tilde{\mathrm{X}}_\psi) \mathrel{+}= \frac{1}{M^a}\log\left(\frac{1}{M^a\,P(\tilde{x}_\phi^{(\tilde{m})})P(\tilde{x}_\psi^{(\tilde{m})})}\right)$    ▷ Compute the addend of the MI between segments for the $\tilde{m}$-th labelling vector of the sensitive AF (Eq. 18).
            if $a \in \mathscr{A}_\phi$ then
                $\mathscr{D}(\mathscr{A}_\phi, \mathscr{A}_\psi; \mathscr{A}_\phi^a) = I(\mathrm{X}_\phi;\mathrm{X}_\psi) - I(\tilde{\mathrm{X}}_\phi;\tilde{\mathrm{X}}_\psi)$    ▷ Compute the *diagnosticity score* (Eq. 24)
                $\mathrm{D}[a] \leftarrow \{a,\ \mathscr{A}_\phi,\ \mathscr{A}_\psi,\ \mathscr{D}(\mathscr{A}_\phi, \mathscr{A}_\psi; \mathscr{A}_\phi^a)\}$    ▷ Append the result to the *diagnosticity vector D*
            if $a \in \mathscr{A}_\psi$ then
                $\mathscr{D}(\mathscr{A}_\phi, \mathscr{A}_\psi; \mathscr{A}_\psi^a) = I(\mathrm{X}_\phi;\mathrm{X}_\psi) - I(\tilde{\mathrm{X}}_\phi;\tilde{\mathrm{X}}_\psi)$    ▷ Compute the *diagnosticity score* (Eq. 24)
                $\mathrm{D}[a] \leftarrow \{a,\ \mathscr{A}_\phi,\ \mathscr{A}_\psi,\ \mathscr{D}(\mathscr{A}_\phi, \mathscr{A}_\psi; \mathscr{A}_\psi^a)\}$    ▷ Append the result to the *diagnosticity vector D*
    return D
```

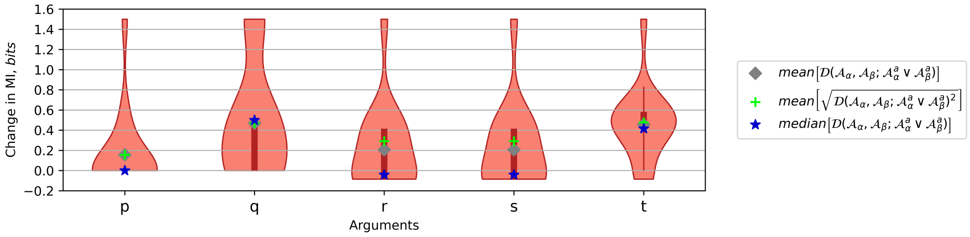


Figure 3: A violin plot showing the distribution of *diagnosticity scores*, as well as the median, mean and average absolute change in MI for the graph $\mathscr{G}$ (Fig. 1a), under complete semantics.

The argument t was arguably the most diagnostic as it had the largest median, mean and average absolute diagnosticity scores (Fig. 3). This result is intuitive because t has symmetric attacks between the arguments r and s such that removing t reduces the number of complete labellings from four (Tab. 1) to three (Tab. 2).

The argument q had the second largest set of diagnosticity scores (Fig. 3). Even though q only attacked one argument (Fig. 1) and its removal does not reduce the number of labellings (Tab. 1), the conditional dependence between t and q meant that q produced greater median and mean diagnosticity scores. An interesting point to note is that the removal q from the initial AF would change the label of the argument p so that it was sceptically labelled IN within the sensitive AF $\mathscr{G}^q$. Thus, for the sensitive AF $\mathscr{G}^q$, the MI scores would equal zero for partitions where p was the only argument in a segment (i.e., $\mathscr{A}^q_\phi = \{p\}$ or $\mathscr{A}^q_\psi = \{p\}$) because the MI is always zero for partitions that only contain arguments which are sceptically labelled.

While the arguments r and s are not the most diagnostic, their removal from the initial AF results in the same distribution of change in MI (Fig. 3), which can be attributed to the symmetry between them in the initial graph. The median change in MI was below zero (Fig. 3) which indicates that, on average, more information was communicated between segments of labelling vectors in the sensitive AFs $\mathscr{G}^r$ and $\mathscr{G}^s$ after the removal of r and s, respectively.

The preliminary result presented in this paper is a first attempt within the literature to combine abstract argumentation and a technique from the information-theoretic literature for sensitivity analysis. We were able to quantify the sensitivity, dependence and robustness of an AF's set of conclusions based upon the arguments it was comprised of.

5. Related Work

The prior literature covers work on probabilistic argumentation, namely the epistemic [7, 8] and constellation [9, 10] approaches. Our method of probability computation differs from both approaches because we do not rely on probability functions to evaluate AFs, as is the case with the epistemic approach. Nor do we iterate through all the permutations of a graph's topology, counting the number of times where a user-chosen set of arguments features in an induced AF's set of extensions, as per the constellation approach. We note one proposal that employed the frequency of individual argument labels to compute marginal probabilities which were then used in combination with a Markov network to semantically evaluate an AF [11]. Our formalised probability calculations (see Eqs. 10, 11, 13 and 14) are the first attempt within the literature to employ partitions of arguments and their labels to compute joint and marginal probabilities.

To the best of our knowledge, there has only been two proposals before this one which use argumentation for sensitivity analysis. The first employs argumentation and Markov random fields to quantify the sensitivity of items of information [12]. The second proposal evaluates the sensitivity of initial weights assigned to arguments within the context of inverse argumentation, where they consider whether changes in an argument's weight affects the acceptability degree of other arguments, computed using gradual semantics [13]. The algorithm presented in this paper combines Dung's original abstract AF and semantics with a set of novel probability calculations to compute the change in MI before and after the removal of an argument of interest.

Aside from the two above works on sensitivity analysis, we consider the argument strength literature to be the most closely related to the DAI. Strength has been represented by arbitrarily assigning a *weight* to an argument [14] or using ranking-based semantics [15, 16, 17, 18, 19]. However, our approach differs from all of the above because we neither choose an argument's weight nor do we not alter the well-known semantics to produce a preorder on the set of arguments.

6. Conclusions and Future Work

The novel algorithm presented in this paper is one of the first attempts within the literature to use of argumentation for sensitivity analysis. We use the labellings output from the evaluation of an AF to compute joint and marginal argument acceptance probabilities which are employed in MI calculations before and after the removal of an argument. We argue that the *diagnosticity* scores output by the DAI provide a holistic quantification of the sensitivity, dependence and robustness of an AF's evaluation. We contend that such a tool would provide benefit to intelligence analysts by algorithmically identifying sensitive arguments found within an analysis.

Future work could involve intelligence analysts, comparing the outputs from the DAI with the outcome from sensitivity analysis conducted by a set of analysts. And, of course, extending the DAI to include more flavoursome AFs is an obvious extension. Another avenue for future work could be to investigate different domains in which the DAI could be applied. For instance, the DAI could aid in decision and deliberation problems, allowing users to utilise the rational logic of argumentation and probability theory to identify and focus on crucial arguments with their task, or by providing reasoners with an indication of which arguments are the most important to attack when analysing a debate. Adding or removing multiple arguments is an interesting idea to pursue as it might better explain the diagnosticity of sets of arguments. Finally, one could argue that the DAI, in its current state, is computationally expensive. Cheaper approaches, such as counting the number of attacks from and to an argument or whether an argument is in a dense area of an AF, should be explored to identify whether they are capable of producing the same results as the DAI.

Acknowledgements

This work was funded by both the Defence Science and Technology Laboratory and the United Kingdom's Engineering and Physical Sciences Research Council (Grant number: EP/S023445/1).

References

[1] R. Heuer, Psychology of Intelligence Analysis, Center for the Study of Intelligence, 1999.

[2] J. Robinson, K. Atkinson, S. Maskell, C. Reed, Identifying diagnostic arguments in abstract argumentation, in: Proc. of COMMA, 2024.

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

[4] M. W. A. Caminada, D. M. Gabbay, A logical account of formal argumentation, Studia Logica 93 (2009) 109–145.

[5] C. E. Shannon, A mathematical theory of communication, The Bell System Technical Journal 27 (1948) 379–423.

[6] T. Cover, J. Thomas, Elements of Information Theory, John Wiley & Sons, 2005, pp. 13–55.

[7] M. Thimm, A probabilistic semantics for abstract argumentation, in: Proc. of ECAI, volume 242, IOS Press, 2012, pp. 750–755.

[8] M. Thimm, A. Hunter, Probabilistic argumentation with incomplete information, in: Frontiers in Artificial Intelligence and Applications, volume 263, 2014, pp. 1033–1034.

[9] P. Dung, P. Thang, Towards (probabilistic) argumentation for jury-based dispute resolution, in: Proc. of COMMA, 2010, pp. 171–182.

[10] H. Li, N. Oren, T. Norman, Probabilistic argumentation frameworks, in: Theories and Applications of Formal Argumentation, volume 7132, 2011, pp. 1–16.

[11] N. Potyka, Abstract argumentation with markov networks, in: Proc. of ECAI, volume 325, IOS Press, 2020, pp. 865–872.

[12] Y. Tang, N. Oren, K. Sycara, Markov argumentation random fields, in: Proc. of AAAI Conference on Artificial Intelligence, 2016.

[13] N. Oren, B. Yun, S. Vesic, M. S. Baptista, Inverse problems for gradual semantics, in: International Joint Conference on Artificial Intelligence, 2022, pp. 2719–2725.

[14] L. Amgoud, C. Cayrol, Integrating preference orderings into argument-based reasoning, in: Qualitative and Quantitative Practical Reasoning, volume 1244, Springer, 1997, pp. 159–170.

[15] C. Cayrol, M.-C. Lagasquie-Schiex, Graduality in argumentation, J. Artif. Intell. Res. (JAIR) 23 (2005) 245–297.

[16] L. Amgoud, J. Ben-Naim, Ranking-based semantics for argumentation frameworks, in: Scalable Uncertainty Management, volume 8078, Springer, 2013, pp. 134–147.

[17] F. Pu, J. Luo, Y. Zhang, G. Luo, Argument ranking with categoriser function, in: Knowledge Science, Engineering and Management, volume 8793, Springer, 2014, pp. 290–301.

[18] L. Amgoud, J. Ben-Naim, D. Doder, S. Vesic, Ranking arguments with compensation-based semantics, in: Proc. of Knowledge Representation and Reasoning, AAAI Press, 2016, p. 12–21.

[19] D. Grossi, S. Modgil, On the graded acceptability of arguments in abstract and instantiated argumentation, Artificial Intelligence 275 (2019) 138–173.