ARTICLE TEMPLATE

A Bayesian Approach to Forward and Inverse Abstract Argumentation Problems

Hiroyuki Kido[a] and Beishui Liao[b]

[a]Cardiff University, Park Place, Cardiff, CF10 3AT, UK; [b]Zhejiang University, 866 Yuhangtang Road, Hangzhou, Zhejiang Province, 310058, P. R. China

ARTICLE HISTORY
Compiled November 2, 2022

ABSTRACT
This paper studies a fundamental mechanism by which conflicts between arguments are drawn from sentiments regarding acceptability of the arguments. Given sets of arguments, an inverse abstract argumentation problem seeks attack relations between arguments such that acceptability semantics interprets each argument in the sets of arguments as being acceptable in each of the attack relations. It is an inverse problem of the traditional problem we refer to as the forward abstract argumentation problem. Given an attack relation, the forward abstract argumentation problem seeks sets of arguments such that acceptability semantics interprets each argument in the sets of arguments as being acceptable in the attack relation. We give a probabilistic model of argumentation-theoretic inference. It is a generative model formalising the process by which acceptability semantics interprets acceptability of arguments in a given attack relation. We show that it gives a broad view of solutions to the forward and inverse abstract argumentation problems. Specifically, solutions to the inverse and forward abstract argumentation problems are shown to be equivalent to a maximum likelihood estimate and maximum likelihood prediction, respectively, which are both available with the generative model. In addition, they are shown to be special cases of the posterior distribution and the evidence, respectively, which are both obtained by probabilistic inference on the generative model. We report an experiment result and application example of the generative model in the inverse problems.

KEYWORDS
Abstract argumentation frameworks; Acceptability semantics; Inverse problems; Generative models; Bayesian statistics; Machine learning

1. Introduction

The world is full of difficult problems. Argumentation is a human cognitive process for the purpose of understanding a lot of these problems. The driving force of argumentation is a conflict of opinions. The ability to detect conflicts is essential for humans to engage in argumentation. The way humans recognise conflicts between arguments can be explained in terms of a simple principle of incompatibility of arguments. Let us take a look at an example to illustrate the principle.

Example 1.1. A professor and a government official argue about a government's policy on the allocation of a research budget.

CONTACT Hiroyuki Kido. Email: KidoH@cardiff.ac.uk

Professor: The government should widely and fairly allocate research funds for diversity of research activities.

Official: We should select and concentrate on promising research in terms of cost effectiveness.

At this point, no logical contradiction, i.e., P and not-P, is explicit from these arguments. The recognition of a logical contradiction thus depends on the context of the arguments and the speakers' or listeners' knowledge.

Suppose that an agent judges these two arguments are incompatible no matter how it interprets them. For example, the agent may think that the wide and fair allocation of research funds and the cost effectiveness cannot be achieved simultaneously. If the agent values the opinion that the diversity of research activities is an essential for persistent progress in science then it may accept only the professor's argument. In contrast, if the agent appreciates short-term profits then it may accept only the official's argument. These kinds of interpretations would contribute to the agent's recognition that these two arguments cause a logical contradiction.

The important observation obtained from this example is that the incompatibility of arguments can be a cause of the recognition of the logical contradiction between the arguments. Conversely, one might think that the recognition of the logical contradiction is an actual cause of the incompatibility of the acceptability. Those who agree with the latter opinion would think that the agent could have made the logical contradiction explicit by revealing unexpressed knowledge for clarification. For example, in Example 1.1, the agent could have made the logical contradiction explicit by saying 'If we select and concentrate on promising research then we cannot widely and fairly allocate research funds.' We think that these two views are two sides of the same coin in human cognition. On the one hand, the incompatibility would help us discover new knowledge that clearly explains the incompatibility. On the other hand, the preexisting but unexpressed knowledge would help us recognise the incompatibility. The prime goal of this paper is to give a formal account of the former process.

There are at least two computational approaches to detect a conflict relation between arguments. The first approach is based on natural language processing or computational linguistics. Given textual discourse, the goal involves identifying individual arguments, their internal structures and their interactions (Lawrence and Reed, 2016; Palau and Moens, 2009). Argumentation mining (Bar-Haim et al., 2017a,b; Boltužić and Šnajder, 2014; Cabrio and Villata, 2013; Lippi and Torroni, 2015, 2016; Mayer et al., 2018; Saint-Dizier, 2018; Toledo-Ronen et al., 2016; Zhao et al., 2017), recognizing textual entailment (Levesque et al., 2012; Silva et al., 2018; Zhao et al., 2017) and natural language inference (Bowman et al., 2015; MacCartney and Manning, 2007) all belong to the first approach. The first approach is successful if a conflict relation obtained with the approach conforms to a human judgement. The second approach is based on the acceptability semantics (Dung, 1995) that emerged from the study of nonmonotonic reasoning (Bench-Capon and Dunne, 2007; Prakken and Vreeswijk, 2001). Given acceptability of arguments, the goal involves identifying an attack relation justifying the acceptability in terms of the acceptability semantics. It is interesting to investigate whether a conflict relation obtained with the second approach is successful in the sense of the goal of the first approach. However, it is not a fundamental requirement of the second approach. The argumentation framework (AF for short) synthesis problem (Niskanen et al., 2016, 2019) based on realizability (Dunne et al., 2015), the abstract structure learning (Riveret and Governatori, 2016), the enforcement in argumentation (Niskanen et al., 2018), the generative model of the abstract argumentation

(Kido, 2018; Kido and Okamoto, 2017), the probabilistic framework for generating instantiated argument graphs (Hunter, 2020) and the explanatory argumentation graph (Riveret, 2020) are all related to the second approach.

The second approach is much less studied compared with the first one, in spite of its importance. Acceptability semantics is a normative theory of human cognition. Thus, a solution to the second approach following the acceptability semantics is normative, meaning that a rational agent ought to accept the solution. Given acceptability of arguments, it tells us which attack relation the agent ought to believe. However, it is more difficult than it seems to find an attack relation in accordance with the acceptability semantics. The acceptability of arguments observed in practice is based on one's sentiments regarding agreement or disagreement on the arguments. It often involves uncertainty, for a variety of reasons such as lack of data and the presence of noise in observation. Thus, what is missing is the way to quantify the uncertainty of solutions so that it follows the acceptability semantics.

Probability theory quantifies uncertainty in a rigorous way. In this paper, we introduce a generative model of abstract argumentation (Dung, 1995). The generative model uses the acceptability semantics and formalises the process by which the acceptability of arguments is generated from the probability distribution over attack relations. Let $\boldsymbol{Att}$ denote an attack relation between arguments and $\boldsymbol{Ext}$ denote a set of sets of arguments that can be accepted simultaneously. The key idea underlying the generative model is to model the right two terms of the following expression so that the value of the left term is exactly or approximately computed.[1]

$$p(\boldsymbol{Att}|\boldsymbol{Ext}) \quad \propto \quad p(\boldsymbol{Ext}|\boldsymbol{Att})p(\boldsymbol{Att})$$

We show that the generative model extends solutions to the following two types of basic problems.

Forward problem Given an attack relation between arguments, the forward problem seeks sets of arguments such that the acceptability semantics interprets the sets of arguments as being acceptable in the attack relation.

Inverse problem Given sets of arguments, the inverse problem seeks an attack relation between arguments such that the acceptability semantics interprets the sets of arguments as being acceptable in the attack relation.

We show that solutions to the inverse and forward problems are respectively equivalent to a maximal likelihood (abbreviated to ML) estimate and ML prediction in the generative model. They are respectively special cases of the posterior distribution and evidence, which are both probabilistic extensions of the solutions to the inverse and forward problems. The posterior distribution is an extension of the solution to the inverse problem in the sense that it deals with noises in observations of sets of arguments. The evidence is an extension of the solution to the forward problem in the sense that it deals with uncertainty of attack relations. We also discuss a compound problem of the inverse and forward problems. Given sets of arguments, the compound problem seeks the acceptability of other sets of arguments by solving the forward problem using the solution to the inverse problem. Finally, we empirically discuss the correctness of the generative model by solving the inverse problems with different noise levels, acceptability semantics, numbers of arguments and restrictions on attack relations.

The main contributions of this paper are as follows. To the best of our knowledge,

[1] $X \propto Y$ denotes 'X is proportional to Y' and thus there is a constant K such that $X = KY$.

the first paper introducing a generative model to the semantics of the abstract argumentation is (Kido and Okamoto, 2017). In a nutshell, it gives problem-independent uncertain argumentation-theoretic inference. It is different from the study of problem-dependent uncertain domain knowledge, e.g., (Bex and Renooij, 2016; Grabmair et al., 2010; Nielsen and Parsons, 2007; Saha and Sen, 2004; Timmer et al., 2015; Vreeswijk, 2005), which is another interesting research direction across argumentation and probability theory. The weakness of the discussions in the papers (Kido, 2018; Kido and Okamoto, 2017), however, is that the generative models introduced by the authors are too specific to discuss what kinds of general problems should be solved and what kinds of problems can be solved based on the idea of generative models. This paper answers the open questions. We for the first time offers a general discussion about a generative model of acceptability of arguments in the context of computational argumentation.

In Section 2, we define the forward and inverse problems of the abstract argumentation as the problems to be solved and discussed in the presence of noise. In Section 3, we give a simple but general generative model for the solutions to the problems. In Sections 4.1 and 4.2, we statistically characterise the properties of the solutions provided by probabilistic inference on the generative model. We then demonstrate its applicability in argumentation mining in Sections 4.3 and 4.4. Section 5 concludes with discussion.

2. Abstract Argumentation Problems

2.1. Forward Problems

An abstract argumentation framework (AF) (Dung, 1995) is a pair $\langle arg, att \rangle$, where arg denotes a set of arguments and att denotes a binary relation on arg. att represents an attack relation between arguments, i.e., $(a, b) \in att$ means 'a attacks b.' Suppose $a \in arg$ and $S \subseteq arg$. S attacks a if, and only if (iff), some member of S attacks a. S is conflict-free iff S attacks none of its members. a is acceptable with respect to S iff S attacks all arguments that attack a. A characteristic function $F : Pow(arg) \rightarrow Pow(arg)$ is defined as $F(S) = \{a | a \text{ is acceptable with respect to } S\}$ where $Pow(arg)$ is the power set of arg. S is admissible iff S is conflict-free and every member of S is acceptable with respect to S. The acceptability semantics (Dung, 1995) defines four types of extensions of AF that intuitively represent sets of acceptable arguments.

- A *preferred extension* is a maximal (with respect to set inclusion) admissible set.
- A conflict-free set S of arguments is a *stable extension* iff S attacks each argument which does not belong to S.
- The *grounded extension* is the least fixed point of F.
- An admissible set S of arguments is a *complete extension* iff each argument, which is acceptable with respect to S, belongs to S.

We call the acceptability semantics for each type of extensions preferred, stable, grounded and complete semantics, and they are denoted by p, s, g and c, respectively. Let ε denote an acceptability semantics, arg a set of arguments, and att an attack relation on arg. When we see ε as a function whose input is an abstract argumentation framework $\langle arg, att \rangle$ and output is the set, denoted by $\boldsymbol{ext}$, of extensions of $\langle arg, att \rangle$ with respect to ε, we can write the relation as follows.

$$\varepsilon(arg, att) = \boldsymbol{ext}$$

We assume that arg and ε are arbitrary but fixed. We define a forward problem of the abstract argumentation, as follows.

Definition 2.1 (Forward problem)**.** Let arg be a set of arguments and ε be an acceptability semantics. Given $att \subseteq Pow(arg \times arg)$, the forward problem of abstract argumentation is to find a set $S \subseteq Pow(arg)$ such that $S = \varepsilon(arg, att)$.

2.2. Inverse Problems

We consider an inverse of the forward problem of the abstract argumentation. In essence, it aims to find an attack relation between arguments from a noisy set of extensions of an abstract argumentation framework. We assume that part of an attack relation, denoted by $\widetilde{att} \subseteq Pow(arg \times arg)$, is known. We now have the following equation.

$$\boldsymbol{ext} = \varepsilon(arg, \widetilde{att} \cup att)$$

We assume that ε, arg and $\widetilde{att}$ are arbitrary but fixed. We define an inverse problem of the abstract argumentation as follows.

Definition 2.2 (Inverse problem)**.** Let arg be a set of arguments, ε be an acceptability semantics and $\widetilde{att} \subseteq Pow(arg \times arg)$ be a known attack relation. Given $S \subseteq Pow(arg)$, the inverse problem of abstract argumentation is to find an attack relation $att \subseteq Pow(arg \times arg)$ such that $S = \varepsilon(arg, \widetilde{att} \cup att)$.

We refer to att as a target attack relation in the inverse problem. A set $S \subseteq Pow(arg)$ is data observed in the inverse problem. It is empirically true that an observation of data often includes some amount of noise. It can be an effect irrelevant to semantics ε or can be a false or inaccurate observation. It may add or remove some arguments from some of the extensions observed, or it may add some new sets of arguments or remove some of the extensions observed. Without loss of generality, we represent a noise η as a subset of $Pow(arg)$, i.e. $\eta \subseteq Pow(arg)$. Let $+$ be a set operator for subtraction and/or addition of elements of the set. The observed data then can be written as follows.

$$S = \varepsilon(arg, \widetilde{att} \cup att) + \eta$$

In other words, what we observe in the inverse problem is a noisy set of extensions where the true set of extensions, i.e., $\varepsilon(arg, \widetilde{att} \cup att)$, and the noise, i.e., η, are inseparably connected. The noisy set of extensions can be the set of extensions of another attack relation. It is thus generally impossible to find a solution to the inverse problem with noise. In reality, we are interested in attack relation att such that the set of extensions, i.e., $\varepsilon(arg, \widetilde{att} \cup att)$, approximates the observation, i.e., S. For the sake of generality, we refrain ourselves from discussing the quality of approximation here. The quality depends on criteria one uses to define the quality. Various criteria are possible, and therefore, they should be discussed in the solution side of the inverse problem.

In Section 4.4, we will demonstrate how the inverse problem can be used to tackle an important task of argumentation mining. From a data science point of view, the inverse problem can be more practical than the forward problem. The input to the inverse problem is in practice a set of sentiments regarding the acceptability of arguments,

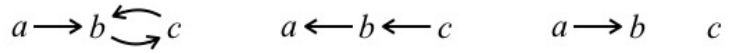
whereas the input to the forward problem is an attack relation between arguments. In contrast to the attack relation that is structured data essentially represented as a graph, the sentiments regarding the acceptability of arguments are unstructured data. The sentiments thus can be collected from the web more easily, e.g., via votes in various social networking services.

2.3. Use of Probability Theory

In this section, we explain why the inverse problem is difficult and why probabilistic approaches are appropriate. The problem, either inverse or direct, is said to be well-posed if a solution exists, the solution is unique if it exists, and the solution depends continuously on the input, i.e., *solution existence*, *solution uniqueness* and *solution stability*, respectively (Aster et al., 2004). We refer to the problem as ill-posed otherwise. Given Definitions 2.1 and 2.2, both the inverse and the forward problems of the abstract argumentation are ill-posed. The solution stability does not hold in the forward or inverse problem as they are not continuous problems but discrete ones. The solution existence and solution uniqueness hold only in the forward problem.

Example 2.3. Let $arg = \{a, b\}$. Given $S = \{\emptyset, \{a, b\}\}$, no solution to the inverse problem exists as there is no attack relation att such that S is the set of extensions of $\langle arg, att\rangle$ with respect to either grounded, preferred, stable or complete semantics.

Example 2.4. Let $arg = \{a, b, c\}$ and $\widetilde{att} = \emptyset$. Given $\boldsymbol{ext} = \{\{a, c\}\}$, the attack relations represented with the directed graphs below are solutions to the inverse problem with respect to grounded, preferred, stable and complete semantics.



The solution uniqueness holds in the inverse problem under certain conditions on an attack relation and semantics.

Proposition 2.5 (Solution uniqueness)**.** *An inverse problem satisfies solution uniqueness if both known and target attack relations are symmetric and irreflexive, and semantics is complete, preferred or stable.*

The restrictions on the attack relations and semantics in Proposition 2.5 can be reasonable in practice. The observation in the inverse problem is a set of sentiments on the acceptability of arguments. The sentiments are generally obtained from various individuals including credulous and skeptical ones. Complete semantics generates various kinds of extensions such as grounded extensions, stable extensions, preferred extensions and complete extensions. The semantics thus suits for capturing the sentiments of both the credulous and skeptical individuals. A reflexive attack relation means the possibility of the presence of a self-attacking argument, which is rare in real argument. A symmetric attack relation means that each attack happens bidirectionally. This is not always true in practice, but even without the direction of attack, information on the presence or absence of attack is useful in many application settings, e.g., argumentation mining.

The fact that a problem is ill-posed does not mean that the problem is defined badly. It implies that the problem is inherently and computationally hard if one seeks alternative solutions from the existence of multiple solutions or approximations from the absence of solutions. Probability theory is useful to handle these situations. There

are at least two other reasons why a probabilistic approach is appropriate. First, almost all of the real arguments are *enthymemes*. An enthymeme is an argument whose premise or conclusion is unexpressed. The existence of an attack relation between enthymemes depends on the contexts of argumentation and the knowledge possessed by the arguers or listeners. Let us take a look at a simple example of the uncertainty of the existence of an attack relation.

Example 2.6. Suppose one hundred individuals are asked to express their opinions about whether there is an attack between the following arguments a and b.

- a: Tweety can fly because it is a bird.
- b: Tweety is a penguin.

Many individuals would find an attack between the arguments, because they know that penguins cannot fly. However, not all individuals would find the attack. Some individuals might not know that penguins cannot fly; some other individuals might know that Tweety is a genetically-altered flying penguin. Now, let us assume that ninety individuals think that there is an attack between the arguments and the remaining ten individuals think that there is no attack. The best we can conclude is that the probability of an attack is 0.9.

Second, the uncertainty of the existence of an attack relation causes the idea of the uncertainty of extensions. A set of extensions would be more probable only when there is a probable attack relation whose set of extensions approximates the set of extensions. Therefore, different attack relations have different influences on the set of extensions observed in the inverse problem. A probabilistic approach allows us to give a formal account of the influence between attack relations and extensions.

3. Abstract Argumentation Model

In this section, a generative model of the abstract argumentation is introduced for solutions to the forward and inverse problems of the abstract argumentation. Using the acceptability semantics, it formalises the probabilistic process of argument-based reasoning by which the acceptability of arguments is generated from the probability distribution over attack relations.

3.1. Probability Distributions

Let arg be a set of arguments. We assume two kinds of random variables. For all $m \in arg \times arg$, Att_m is a random variable representing the truth of whether there is an attack from its left to right elements of m. att_m represents a value of Att_m, either 0 or 1 meaning false or true, respectively.[2] For all $d \subseteq arg$, Ext_d is a random variable representing the truth of whether d is an extension. ext_d represents a value of Ext_d, either 0 or 1. $\boldsymbol{Att}$ denotes a sequence of Att_m, and $\boldsymbol{att}$ a sequence of att_m. Similarly, $\boldsymbol{Ext}$ denotes a sequence of Ext_d, and $\boldsymbol{ext}$ a sequence of ext_d. $\boldsymbol{att}$ and $\boldsymbol{ext}$ thus correspond respectively to an attack relation and extensions in the abstract argumentation (Dung, 1995). We do not distinguish them unless otherwise noted.

Assuming that arg and acceptability semantics are arbitrary but fixed, we study

[2]For the sake of simplicity, m also represents a set of two arguments. Att_m in this case represents the existence of a symmetric attack relation between the arguments in m.

the relation between $\boldsymbol{Att}$ and $\boldsymbol{Ext}$. It is natural to assume that if the probability that a attacks b is $\lambda_{(a,b)}$ then the probability that a does not attack b is $1 - \lambda_{(a,b)}$. We thus define the probability distribution over an attack-relation variable as follows.

Definition 3.1 (Attack distribution)**.** Let Att_m be a random variable of an attack relation and λ_m be a constant such that $0 \leq \lambda_m \leq 1$. The probability distribution over Att_m, denoted by $p(Att_m)$, is given by

$$p(Att_m) = \lambda_m^{Att_m}(1 - \lambda_m)^{1-Att_m}.$$

It is obvious that $p(Att_m) = \lambda_m$ if $Att_m = 1$ and $p(Att_m) = 1 - \lambda_m$ if $Att_m = 0$. The distribution is called a *Bernoulli distribution* (Uspensky, 1937), often used to represent a discrete probability distribution with binary values. One can assume the uniform distribution, i.e., $\lambda_m = 0.5$, when no knowledge on the presence of the attack relation in m is available or assumed. One can alternatively apply the result of textual analysis with natural language processing to give a value to each λ_m. We leave the values of λ_m unspecified to make our idea open to various application scenarios. Note that, as will be illustrated in Example 3.11, λ_m is updated in accordance with data. Collecting data is thus more important than trying to accurately specify the values of λ_m in our Bayesian approach.

We next consider how to handle noisy sets of extensions given an attack relation. The acceptability semantics (Dung, 1995) defines ext_d given $\boldsymbol{att}$. Let us assume a constant $\theta_{d|\boldsymbol{att}}$ ($0 \leq \theta_{d|\boldsymbol{att}} \leq 1$) representing the probability that d is an extension of the abstract argumentation framework given by attack relation $\boldsymbol{att}$. The probability distribution over Ext_d given $\boldsymbol{att}$ can be defined using a Bernoulli distribution with parameter $\theta_{d|\boldsymbol{att}}$.

Definition 3.2 (Extension distribution)**.** Let Ext_d be an extension variable, $\boldsymbol{att}$ be a sequence of values of attack-relation variables and $\theta_{d|\boldsymbol{att}}$ be a constant such that $0 \leq \theta_{d|\boldsymbol{att}} \leq 1$. The probability distribution over Ext_d given $\boldsymbol{att}$, denoted by $p(Ext_d|\boldsymbol{att})$, is given by

$$p(Ext_d|\boldsymbol{att}) = \theta_{d|\boldsymbol{att}}^{Ext_d}(1 - \theta_{d|\boldsymbol{att}})^{1-Ext_d}.$$

The constant $\theta_{d|att}$ is the parameter of the extension distribution. In the next section, we will present three different instantiations for the constant $\theta_{d|att}$.

3.2. Likelihoods

This section defines a deterministic parameter, linear parameter and exponential parameter for the constant $\theta_{d|att}$. We introduce the first two parameters as straightforward applications of the acceptability semantics, and the third parameter as a generalisation of the first two parameters. In Section 4, we will conduct experiments to discuss the correctness of the third parameter.

From the perspective of the acceptability semantics, the constant $\theta_{d|\boldsymbol{att}}$ should have a high value when there is an extension e of the abstract argumentation framework $\langle arg, \boldsymbol{att}\rangle$ such that e is close to d. The closeness can be measured by the similarity,

denoted by $sim(e, d)$, defined as follows[3].

$$sim(e, d) = |\{a \in arg | a \in e, a \in d\} \cup \{a \in arg | a \notin e, a \notin d\}|$$

We define a deterministic parameter as follows.

Definition 3.3 (Deterministic parameter)**.** Let $d \subseteq arg$ and $\boldsymbol{att}$ be an attack relation on arg. $\theta_{d|\boldsymbol{att}} \in [0, 1]$ is a deterministic parameter if it is given by

$$\theta_{d|\boldsymbol{att}} = \begin{cases} 1 & d \in \varepsilon(arg, \boldsymbol{att}) \\ 0 & otherwise. \end{cases}$$

Therefore, $\theta_{d|\boldsymbol{att}} = 1$ holds iff d is an extension of the argumentation framework of $\boldsymbol{att}$.

The second parameter is a linear parameter. We define $\theta_{d|\boldsymbol{att}}$ using the linear function $f(x) = x/|arg|$ with respect to x, where x is the maximum similarity between d and an extension of the argumentation framework of $\boldsymbol{att}$.

Definition 3.4 (Linear parameter)**.** Let $d \subseteq arg$ and $\boldsymbol{att}$ be an attack relation on arg. $\theta_{d|\boldsymbol{att}} \in [0, 1]$ is a linear parameter if it is given by

$$\theta_{d|\boldsymbol{att}} = \frac{1}{|arg|} \max \left\{ sim(d, e) \;\middle|\; e \in \varepsilon(arg, \boldsymbol{att}) \right\}.$$

$\theta_{d|\boldsymbol{att}}$ thus increases with $1/|arg|$ when another argument is in common between e and d.

The third parameter is an exponential parameter. We define $\theta_{d|\boldsymbol{att}}$ using the exponential function $f(x) = w^x/w^{|arg|}$ with respect to x, where w is a constant and x is the maximum similarity between d and an extension of the argumentation framework of $\boldsymbol{att}$. The following definition uses the normalisation of $f(x)$ where its domain and range are $[0, |arg|]$ and $[0, 1]$, respectively.

Definition 3.5 (Exponential parameter)**.** Let $d \subseteq arg$, $\boldsymbol{att}$ be an attack relation on arg and $w > 1$. $\theta_{d|\boldsymbol{att}} \in [0, 1]$ is an exponential parameter if it is given by

$$\theta_{d|\boldsymbol{att}} = \frac{1}{w^{|arg|} - 1} \max \left\{ w^{sim(e,d)} - 1 \;\middle|\; e \in \varepsilon(arg, \boldsymbol{att}) \right\}.$$

Given a large value w, $\theta_{d|\boldsymbol{att}}$ approximates $w^{sim(e,d)}/w^{|arg|}$. In this case, $\theta_{d|\boldsymbol{att}}$ increases w times when another argument is in common between e and d.

The linear and exponential parameters are exactly the results of the normalisations of a linear function and an exponential function, respectively. The exponential parameter has good properties. First, the deterministic parameter is a special case of the exponential parameter.

Proposition 3.6. *Let $\theta_{d|\boldsymbol{att}}$ be an exponential parameter and $\phi_{d|\boldsymbol{att}}$ be a deterministic parameter.* $\lim_{w\to\infty} \theta_{d|\boldsymbol{att}} = \phi_{d|\boldsymbol{att}}$ *holds.*

Second, the linear parameter is also a special case of the exponential parameter.

[3] $|X|$ denotes the cardinality of set X.

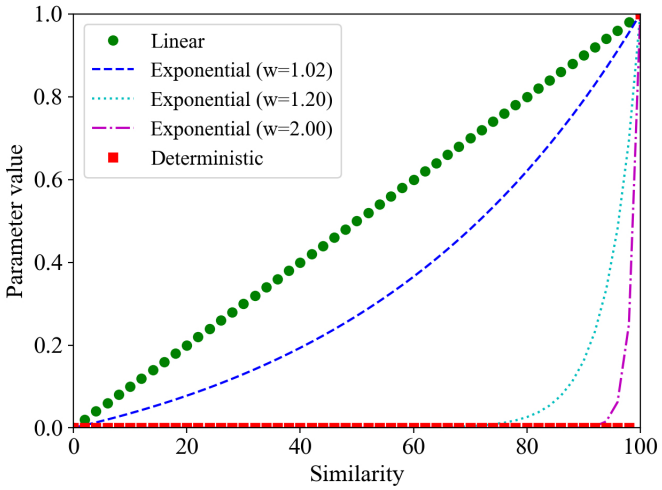


Figure 1. The horizontal axis shows the similarity between an observation and the best extension given 100 arguments. The vertical axis shows parameter values.

Proposition 3.7. *Let* $\theta_{d|\boldsymbol{att}}$ *be an exponential parameter and* $\phi_{d|\boldsymbol{att}}$ *be a linear parameter.* $\lim_{w\to 1} \theta_{d|\boldsymbol{att}} = \phi_{d|\boldsymbol{att}}$ *holds.*

Figure 1 shows that both of the deterministic and linear parameters are extreme cases of the exponential parameter.

Example 3.8. Table 1 shows all possible linear and exponential parameters given $arg = \{a, b, c\}$ and $\boldsymbol{att} = (att_{\{a,b\}}, att_{\{a,c\}}, att_{\{b,c\}})$.

Example 3.8 shows that the exponential parameters give a relatively sharp distribution compared with the linear parameters. As more arguments are given, the linear parameters give a relatively flat distribution, which in practice makes it difficult to find exact or approximate solutions to the inverse abstract argumentation problem. For the sake of generality and practicality, we assume the exponential parameter unless otherwise stated.

3.3. Graphical Model

Figure 2 shows a graphical model of the dependency of the random variables and parameters we introduced in Section 3.1. Each open white-circle represents a random variable, each filled black-circle represents a parameter given in prior to the use of the model, and each rectangle, often referred to as a plate, represents that all the circles on the plate are duplicated the number of times specified in the right bottom corner of the plate. M, L and D denote the numbers of attack-relation variables, attack relations and extensions, respectively. There are thus M random variables of Att_m and M parameters of λ_m, for each pair m of arguments, D random variables of Ext_d, for each set d of arguments, and $L \times D$ parameters of $\theta_{d|\boldsymbol{att}}$, for each pair of d and attack relation $\boldsymbol{att}$. Each arrow represents the inference causality of the abstract argumentation. The parameter λ_m causes the (prior) probability distribution over Att_m. The sequence, i.e., $\boldsymbol{Att}$, of Att_m and the parameter $\theta_{d|\boldsymbol{att}}$ cause the probability distribution over Ext_d.

Table 1. Linear parameters (upper) and exponential parameters (lower) defined with complete semantics. θ_X is an abbreviation for $\theta_{X|\boldsymbol{att}}$ where $\boldsymbol{att} = (att_{\{a,b\}}, att_{\{a,c\}}, att_{\{b,c\}})$.

$\boldsymbol{att}$	$\theta_\emptyset$	$\theta_{\{a\}}$	$\theta_{\{b\}}$	$\theta_{\{c\}}$	$\theta_{\{a,b\}}$	$\theta_{\{a,c\}}$	$\theta_{\{b,c\}}$	$\theta_{\{a,b,c\}}$
$(0,0,0)$	0	$1/3$	$1/3$	$1/3$	$2/3$	$2/3$	$2/3$	1
$(1,0,0)$	$2/3$	$2/3$	$2/3$	1	$1/3$	1	1	$2/3$
$(0,1,0)$	$2/3$	$2/3$	1	$2/3$	1	$1/3$	1	$2/3$
$(1,1,0)$	1	1	$2/3$	$2/3$	$2/3$	$2/3$	1	$2/3$
$(0,0,1)$	$2/3$	1	$2/3$	$2/3$	1	1	$1/3$	$2/3$
$(1,0,1)$	1	$2/3$	1	$2/3$	$2/3$	1	$2/3$	$2/3$
$(0,1,1)$	1	$2/3$	$2/3$	1	1	$2/3$	$2/3$	$2/3$
$(1,1,1)$	1	1	1	1	$2/3$	$2/3$	$2/3$	$1/3$
$(0,0,0)$	0	$\frac{w-1}{w^3-1}$	$\frac{w-1}{w^3-1}$	$\frac{w-1}{w^3-1}$	$\frac{w^2-1}{w^3-1}$	$\frac{w^2-1}{w^3-1}$	$\frac{w^2-1}{w^3-1}$	1
$(1,0,0)$	$\frac{w^2-1}{w^3-1}$	$\frac{w^2-1}{w^3-1}$	$\frac{w^2-1}{w^3-1}$	1	$\frac{w-1}{w^3-1}$	1	1	$\frac{w^2-1}{w^3-1}$
$(0,1,0)$	$\frac{w^2-1}{w^3-1}$	$\frac{w^2-1}{w^3-1}$	1	$\frac{w^2-1}{w^3-1}$	1	$\frac{w-1}{w^3-1}$	1	$\frac{w^2-1}{w^3-1}$
$(1,1,0)$	1	1	$\frac{w^2-1}{w^3-1}$	$\frac{w^2-1}{w^3-1}$	$\frac{w^2-1}{w^3-1}$	$\frac{w^2-1}{w^3-1}$	1	$\frac{w^2-1}{w^3-1}$
$(0,0,1)$	$\frac{w^2-1}{w^3-1}$	1	$\frac{w^2-1}{w^3-1}$	$\frac{w^2-1}{w^3-1}$	1	1	$\frac{w-1}{w^3-1}$	$\frac{w^2-1}{w^3-1}$
$(1,0,1)$	1	$\frac{w^2-1}{w^3-1}$	1	$\frac{w^2-1}{w^3-1}$	$\frac{w^2-1}{w^3-1}$	1	$\frac{w^2-1}{w^3-1}$	$\frac{w^2-1}{w^3-1}$
$(0,1,1)$	1	$\frac{w^2-1}{w^3-1}$	$\frac{w^2-1}{w^3-1}$	1	1	$\frac{w^2-1}{w^3-1}$	$\frac{w^2-1}{w^3-1}$	$\frac{w^2-1}{w^3-1}$
$(1,1,1)$	1	1	1	1	$\frac{w^2-1}{w^3-1}$	$\frac{w^2-1}{w^3-1}$	$\frac{w^2-1}{w^3-1}$	$\frac{w-1}{w^3-1}$

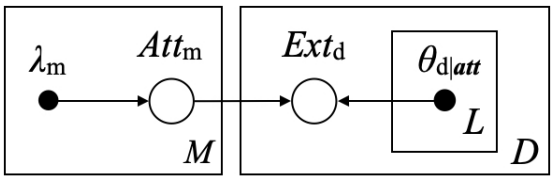


Figure 2. Dependency of the elements of the abstract argumentation model.

Example 3.9. Figure 3 shows an example of Figure 2. Given two arguments a and b, it expresses the dependency without the plate notation.

So far, we defined the prior distribution $p(Att_m)$ as a Bernoulli distribution with the parameter λ_m, and the likelihood distribution $p(Ext_d|\boldsymbol{att})$ as a Bernoulli distribution with the parameter $\theta_{d|\boldsymbol{att}}$. Given all of the parameters, they give the full joint distribution over all of the random variables. We thus call $\{p(\boldsymbol{Att}), p(\boldsymbol{Ext}|\boldsymbol{Att})\}$ an abstract argumentation model, denoted by $\mathcal{M}$. When the parameters need to be specified, we write it as $\{p(\boldsymbol{Att}|\boldsymbol{\lambda}), p(\boldsymbol{Ext}|\boldsymbol{Att}, \boldsymbol{\theta})\}$ where $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ are sequences of λ_m and $\theta_{d|\boldsymbol{att}}$, respectively.

Example 3.10. Suppose that $\boldsymbol{Att}$, $\boldsymbol{Ext}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ are given as follows.

- $\boldsymbol{Att} = (Att_{(a,b)}, Att_{(b,a)})$
- $\boldsymbol{Ext} = (Ext_\emptyset, Ext_{\{a\}}, Ext_{\{b\}}, Ext_{\{a,b\}})$
- $\boldsymbol{\lambda} = (\lambda_{(a,b)}, \lambda_{(b,a)})$
- $\boldsymbol{\theta} = (\theta_{d|(att_{(a,b)}, att_{(b,a)})} | d \in \{\emptyset, \{a\}, \{b\}, \{a,b\}\}, att_{(a,b)}, att_{(b,a)} \in \{0,1\})$

$\{p(\boldsymbol{Att}|\boldsymbol{\lambda}), p(\boldsymbol{Ext}|\boldsymbol{Att}, \boldsymbol{\theta})\}$ is the abstract argumentation model shown in Figure 3.

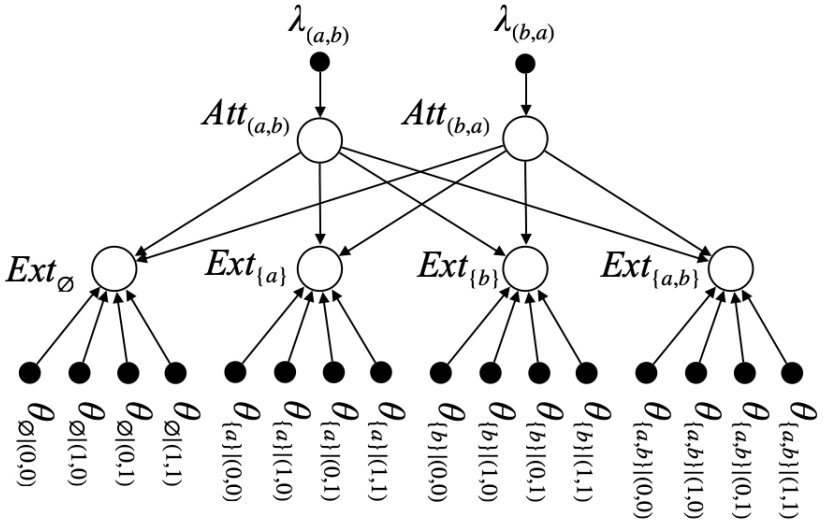


Figure 3. Example of Figure 2 without a plate notation.

3.4. Probabilistic Inference of Attack Relations

The abstract argumentation model $\mathcal{M}$ is a *generative model* as it captures the argumentation-theoretic inference. It tells us how an extension is generated from an attack relation in accordance with the acceptability semantics. In an inverse problem, we use the model to trace the dependency back to the argumentation framework from given extensions. Technically speaking, this is executed by calculating the posterior distribution over attack relations given extensions. Using Bayes' theorem, we have

$$\begin{aligned} p(\boldsymbol{Att}|\boldsymbol{ext}) &= \frac{p(\boldsymbol{ext}|\boldsymbol{Att})p(\boldsymbol{Att})}{p(\boldsymbol{ext})} \propto p(\boldsymbol{ext}|\boldsymbol{Att})p(\boldsymbol{Att}) \\ &= \prod_d p(ext_d|\boldsymbol{Att}) \prod_m p(Att_m) \\ &= \prod_d \theta_{d|\boldsymbol{Att}}^{ext_d}(1-\theta_{d|\boldsymbol{Att}})^{1-ext_d} \prod_m \lambda_m^{Att_m}(1-\lambda_m)^{1-Att_m} \end{aligned}$$

where $x \propto y$ means 'x is proportional to y' and thus there is a constant K such that $x = Ky$. The derivation follows the two assumptions of the abstract argumentation model. Firstly, as shown in Figure 2, attacks between arguments are assumed to be independent, i.e., $p(att_{m_1}, att_{m_2}) = p(att_{m_1})p(att_{m_2})$, for pairs m_1 and m_2 of arguments. We use this assumption to implement approximate inference following our discussion in Section 4. Note that this assumption comes from the fact that in abstract argumentation frameworks we should have no prior information about the presence or absence of attacks between two arguments. This is, however, not often the case in other argumentation frameworks with internal structures of arguments. Secondly, each extension is assumed to be independently distributed from the same distribution over attack relations. In other words, two extensions ext_{d_1} and ext_{d_2} are conditionally independent given an attack relation $\boldsymbol{att}$, i.e., $p(ext_{d_1}, ext_{d_2}|\boldsymbol{att}) = p(ext_{d_1}|\boldsymbol{att})p(ext_{d_2}|\boldsymbol{att})$. It is not an assumption introduced in this paper, but is a property of the abstract argumentation. Indeed, the conditional independence can also be equivalently written as

$p(ext_{d_1}|ext_{d_2}, \boldsymbol{att}) = p(ext_{d_1}|\boldsymbol{att})$. It is the case in the abstract argumentation because only an attack relation affects extensions. It results in the desirable property that the posterior distribution over attack relations can be updated successively whenever an extension is observed. In fact, it is obvious from the above equation that we have the following equation.

$$p(\boldsymbol{Att}|\boldsymbol{ext}) \quad \propto \quad p(\boldsymbol{Att}) \prod_d p(ext_d|\boldsymbol{Att})$$

When another set e of acceptable arguments is observed, the above equation leads to

$$\begin{aligned} p(\boldsymbol{Att}|\boldsymbol{ext}, ext_e) &\propto p(\boldsymbol{Att}) p(ext_e|\boldsymbol{Att}) \prod_d p(ext_d|\boldsymbol{Att}) \\ &\propto p(\boldsymbol{Att}|\boldsymbol{ext}) p(ext_e|\boldsymbol{Att}). \end{aligned}$$

Therefore, the most recent posterior distribution is proportional to the product of the previous posterior distribution and the likelihood of the new observation.

Example 3.11. We here see how the probability distribution over attack relations is updated. Let us assume set $arg = \{a, b, c\}$ of arguments and three random variables $Att_{\{a,b\}}$, $Att_{\{a,c\}}$, $Att_{\{b,c\}}$ of an attack relation. Now, suppose that we observe $\{a\}$, i.e., $Ext_{\{a\}} = 1$. The posterior distribution over attack-relation variables given the observation is given by

$$\begin{aligned} & p(Att_{\{a,b\}}, Att_{\{a,c\}}, Att_{\{b,c\}}|Ext_{\{a\}} = 1) \\ \propto \; & p(Ext_{\{a\}} = 1|Att_{\{a,b\}}, Att_{\{a,c\}}, Att_{\{b,c\}}) p(Att_{\{a,b\}}) p(Att_{\{a,c\}}) p(Att_{\{b,c\}}) \\ = \; & \theta_{\{a\}|Att_{\{a,b\}}, Att_{\{a,c\}}, Att_{\{b,c\}}} \prod_{m \in \{\{a,b\},\{a,c\},\{b,c\}\}} \lambda_m^{Att_m} (1 - \lambda_m)^{1 - Att_m}. \end{aligned}$$

Let $\boldsymbol{ext}$ be $(Ext_{\{a\}} = 1)$ and $\theta_{d|\boldsymbol{att}}$ be the exponential parameter shown in Table 1 with $w = 2$. The posterior distribution is then given as follows.

$$\begin{aligned} p(0,0,0|\boldsymbol{ext}) &\propto (1/7)(1 - \lambda_{\{a,b\}})(1 - \lambda_{\{a,c\}})(1 - \lambda_{\{b,c\}}) \\ p(1,0,0|\boldsymbol{ext}) &\propto (3/7)\lambda_{\{a,b\}}(1 - \lambda_{\{a,c\}})(1 - \lambda_{\{b,c\}}) \\ p(0,1,0|\boldsymbol{ext}) &\propto (3/7)(1 - \lambda_{\{a,b\}})\lambda_{\{a,c\}}(1 - \lambda_{\{b,c\}}) \\ p(1,1,0|\boldsymbol{ext}) &\propto \lambda_{\{a,b\}}\lambda_{\{a,c\}}(1 - \lambda_{\{b,c\}}) \\ p(0,0,1|\boldsymbol{ext}) &\propto (1 - \lambda_{\{a,b\}})(1 - \lambda_{\{a,c\}})\lambda_{\{b,c\}} \\ p(1,0,1|\boldsymbol{ext}) &\propto (3/7)\lambda_{\{a,b\}}(1 - \lambda_{\{a,c\}})\lambda_{\{b,c\}} \\ p(0,1,1|\boldsymbol{ext}) &\propto (3/7)(1 - \lambda_{\{a,b\}})\lambda_{\{a,c\}}\lambda_{\{b,c\}} \\ p(1,1,1|\boldsymbol{ext}) &\propto \lambda_{\{a,b\}}\lambda_{\{a,c\}}\lambda_{\{b,c\}} \end{aligned}$$

Next, we suppose another observation $\{b\}$, i.e., $Ext_{\{b\}} = 1$. The posterior distribution is updated as follows.

$$\begin{aligned} & p(Att_{\{a,b\}}, Att_{\{a,c\}}, Att_{\{b,c\}}|Ext_{\{a\}} = 1, Ext_{\{b\}} = 1) \\ \propto \; & p(Att_{\{a,b\}}, Att_{\{a,c\}}, Att_{\{b,c\}}|Ext_{\{a\}} = 1) p(Ext_{\{b\}} = 1|Att_{\{a,b\}}, Att_{\{a,c\}}, Att_{\{b,c\}}) \\ \propto \; & p(Att_{\{a,b\}}, Att_{\{a,c\}}, Att_{\{b,c\}}|Ext_{\{a\}} = 1) \theta_{\{b\}|Att_{\{a,b\}}, Att_{\{a,c\}}, Att_{\{b,c\}}} \end{aligned}$$

Now, let $\boldsymbol{ext}$ be $(Ext_{\{a\}} = 1, Ext_{\{b\}} = 1)$. The posterior distribution is updated as follows.

$$\begin{aligned}
p(0,0,0|\boldsymbol{ext}) &\propto (1/7)^2(1-\lambda_{\{a,b\}})(1-\lambda_{\{a,c\}})(1-\lambda_{\{b,c\}}) \\
p(1,0,0|\boldsymbol{ext}) &\propto (3/7)^2\lambda_{\{a,b\}}(1-\lambda_{\{a,c\}})(1-\lambda_{\{b,c\}}) \\
p(0,1,0|\boldsymbol{ext}) &\propto (3/7)(1-\lambda_{\{a,b\}})\lambda_{\{a,c\}}(1-\lambda_{\{b,c\}}) \\
p(1,1,0|\boldsymbol{ext}) &\propto (3/7)\lambda_{\{a,b\}}\lambda_{\{a,c\}}(1-\lambda_{\{b,c\}}) \\
p(0,0,1|\boldsymbol{ext}) &\propto (3/7)(1-\lambda_{\{a,b\}})(1-\lambda_{\{a,c\}})\lambda_{\{b,c\}} \\
p(1,0,1|\boldsymbol{ext}) &\propto (3/7)\lambda_{\{a,b\}}(1-\lambda_{\{a,c\}})\lambda_{\{b,c\}} \\
p(0,1,1|\boldsymbol{ext}) &\propto (3/7)^2(1-\lambda_{\{a,b\}})\lambda_{\{a,c\}}\lambda_{\{b,c\}} \\
p(1,1,1|\boldsymbol{ext}) &\propto \lambda_{\{a,b\}}\lambda_{\{a,c\}}\lambda_{\{b,c\}}
\end{aligned}$$

We further suppose an additional observation $\{c\}$, i.e., $Ext_{\{c\}} = 1$. The posterior distribution is updated as follows.

$$\begin{aligned}
& p(Att_{\{a,b\}}, Att_{\{a,c\}}, Att_{\{b,c\}}|Ext_{\{a\}} = 1, Ext_{\{b\}} = 1, Ext_{\{c\}} = 1) \\
\propto\ & p(Att_{\{a,b\}}, Att_{\{a,c\}}, Att_{\{b,c\}}|Ext_{\{a\}} = 1, Ext_{\{b\}} = 1)\theta_{\{c\}|Att_{\{a,b\}},Att_{\{a,c\}},Att_{\{b,c\}}}
\end{aligned}$$

Let $\boldsymbol{ext}$ be $(Ext_{\{a\}} = 1, Ext_{\{b\}} = 1, Ext_{\{c\}} = 1)$. The posterior distribution is updated as follows.

$$\begin{aligned}
p(0,0,0|\boldsymbol{ext}) &\propto (1/7)^3(1-\lambda_{\{a,b\}})(1-\lambda_{\{a,c\}})(1-\lambda_{\{b,c\}}) \\
p(1,0,0|\boldsymbol{ext}) &\propto (3/7)^2\lambda_{\{a,b\}}(1-\lambda_{\{a,c\}})(1-\lambda_{\{b,c\}}) \\
p(0,1,0|\boldsymbol{ext}) &\propto (3/7)^2(1-\lambda_{\{a,b\}})\lambda_{\{a,c\}}(1-\lambda_{\{b,c\}}) \\
p(1,1,0|\boldsymbol{ext}) &\propto (3/7)^2\lambda_{\{a,b\}}\lambda_{\{a,c\}}(1-\lambda_{\{b,c\}}) \\
p(0,0,1|\boldsymbol{ext}) &\propto (3/7)^2(1-\lambda_{\{a,b\}})(1-\lambda_{\{a,c\}})\lambda_{\{b,c\}} \\
p(1,0,1|\boldsymbol{ext}) &\propto (3/7)^2\lambda_{\{a,b\}}(1-\lambda_{\{a,c\}})\lambda_{\{b,c\}} \\
p(0,1,1|\boldsymbol{ext}) &\propto (3/7)^2(1-\lambda_{\{a,b\}})\lambda_{\{a,c\}}\lambda_{\{b,c\}} \\
p(1,1,1|\boldsymbol{ext}) &\propto \lambda_{\{a,b\}}\lambda_{\{a,c\}}\lambda_{\{b,c\}}
\end{aligned}$$

In general, we suppose that $Ext_{\{a\}} = 1$, $Ext_{\{b\}} = 1$ and $Ext_{\{c\}} = 1$ are repeatedly observed N times in total in this order. Given N observations, denoted by $\boldsymbol{ext}$, the posterior distribution is given as follows.

$$\begin{aligned}
p(0,0,0|\boldsymbol{ext}) &\propto (1/7)^N(1-\lambda_{\{a,b\}})(1-\lambda_{\{a,c\}})(1-\lambda_{\{b,c\}}) \\
p(1,0,0|\boldsymbol{ext}) &\propto (3/7)^{N-\lfloor N/3 \rfloor}\lambda_{\{a,b\}}(1-\lambda_{\{a,c\}})(1-\lambda_{\{b,c\}}) \\
p(0,1,0|\boldsymbol{ext}) &\propto (3/7)^{N-\lfloor (N+1)/3 \rfloor}(1-\lambda_{\{a,b\}})\lambda_{\{a,c\}}(1-\lambda_{\{b,c\}}) \\
p(1,1,0|\boldsymbol{ext}) &\propto (3/7)^{N-\lfloor (N+2)/3 \rfloor}\lambda_{\{a,b\}}\lambda_{\{a,c\}}(1-\lambda_{\{b,c\}}) \\
p(0,0,1|\boldsymbol{ext}) &\propto (3/7)^{N-\lfloor (N+2)/3 \rfloor}(1-\lambda_{\{a,b\}})(1-\lambda_{\{a,c\}})\lambda_{\{b,c\}} \\
p(1,0,1|\boldsymbol{ext}) &\propto (3/7)^{N-\lfloor (N+1)/3 \rfloor}\lambda_{\{a,b\}}(1-\lambda_{\{a,c\}})\lambda_{\{b,c\}} \\
p(0,1,1|\boldsymbol{ext}) &\propto (3/7)^{N-\lfloor N/3 \rfloor}(1-\lambda_{\{a,b\}})\lambda_{\{a,c\}}\lambda_{\{b,c\}} \\
p(1,1,1|\boldsymbol{ext}) &\propto \lambda_{\{a,b\}}\lambda_{\{a,c\}}\lambda_{\{b,c\}}
\end{aligned}$$

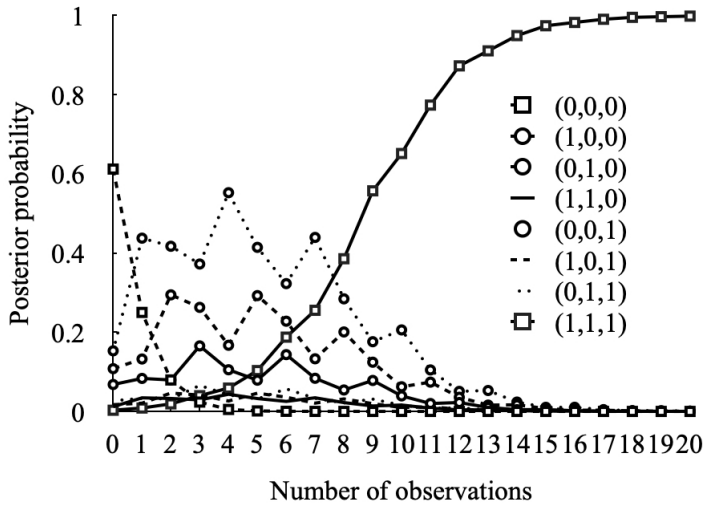


Figure 4. The horizontal axis shows the twenty iterated observations of $Ext_{\{a\}} = 1$, $Ext_{\{b\}} = 1$ and $Ext_{\{c\}} = 1$ in this order. The vertical axis shows the posterior probability of each attack relation $(att_{\{a,b\}}, att_{\{a,c\}}, att_{\{b,c\}})$. We assumed the following parameters: $\lambda_{\{a,b\}} = 0.1$, $\lambda_{\{a,c\}} = 0.15$, $\lambda_{\{b,c\}} = 0.2$ and the exponential parameters $\theta_{d|\boldsymbol{att}}$ shown in Table 1 with $w = 2$.

Here, $\lfloor x \rfloor$ denotes the floor function that returns the maximum integer that is equal to or less than x. Figure 4 shows the posterior distribution over attack relations versus the number of observations. When the number of observations increases, it is observed that the posterior probability of the attack relation, $(Att_{\{a,b\}} = 1, Att_{\{a,c\}} = 1, Att_{\{b,c\}} = 1)$, converges to one. The result is reasonable in terms of complete semantics because the attack relation successfully explained all of the observations.

3.5. Approximate Inference Algorithm

In practice, we can use the model $\mathcal{M}$ not only for solving the inverse and forward problems, but also for solving their compound problems. Given noisy data about sets of acceptable arguments, the compound problem is to find other sets of acceptable arguments. Here, the output sets are predicted using the attack relations estimated from the input sets. In a nutshell, it solves the forward problem using the result of the inverse problem. Concretely, the compound problem is based on

$$\begin{aligned} p(Ext_e|\boldsymbol{ext}) &= \sum_{\boldsymbol{att}} p(Ext_e|\boldsymbol{att}) p(\boldsymbol{att}|\boldsymbol{ext}) \\ &= \sum_{\boldsymbol{att}} \theta_{e|\boldsymbol{att}}^{Ext_e} (1 - \theta_{e|\boldsymbol{att}})^{1-Ext_e} p(\boldsymbol{att}|\boldsymbol{ext}). \end{aligned}$$

However, the problem associated with the compound problem (and also the inverse problem), is that the posterior probability $p(\boldsymbol{att}|\boldsymbol{ext})$ is generally not exactly calculable, due to its high complexity. We thus calculate its approximation, denoted by $\hat{p}(\boldsymbol{att}|\boldsymbol{ext})$. We obtain the approximation using Gibbs sampling (Geman and Geman, 1984), which is a simple and widely applicable Markov chain Monte Carlo algorithm. It repeatedly updates a value of each random variable, one by one, using its posterior distribution given values of all remaining random variables. For example, in the $(i+1)$-

th iteration of the Gibbs sampling procedure, the presence of an attack, denoted by att_m^{i+1}, between two arguments labeled m $(1 \leq m \leq M)$ is sampled as follows.

$$\begin{aligned} att_1^{i+1} &\sim p(Att_1|att_2^i, att_3^i, \cdots, att_M^i, \boldsymbol{ext}) \\ att_2^{i+1} &\sim p(Att_2|att_1^{i+1}, att_3^i, \cdots, att_M^i, \boldsymbol{ext}) \\ &\vdots \\ att_M^{i+1} &\sim p(Att_M|att_1^{i+1}, att_2^{i+1}, \cdots, att_{M-1}^{i+1}, \boldsymbol{ext}) \end{aligned}$$

Here, $\sim$ denotes that the left value is sampled from the right distribution. Let $\boldsymbol{att}_{\backslash m}^{(i+1)}$ denote all attack-relation values except att_m in the $(i+1)$-th iteration, i.e., $\boldsymbol{att}_{\backslash m}^{(i+1)} = (att_1^{i+1}, att_2^{i+1}, \cdots, att_{m-1}^{i+1}, att_{m+1}^i, \cdots, att_M^i)$. The right terms of the above expressions can be written as the following single expression.

$$\begin{aligned} p(Att_m|\boldsymbol{att}_{\backslash m}^{(i+1)}, \boldsymbol{ext}) &= \frac{p(\boldsymbol{ext}|Att_m, \boldsymbol{att}_{\backslash m}^{(i+1)})p(Att_m)p(\boldsymbol{att}_{\backslash m}^{(i+1)})}{p(\boldsymbol{att}_{\backslash m}^{(i+1)}, \boldsymbol{ext})} \\ \propto\ & p(\boldsymbol{ext}|Att_m, \boldsymbol{att}_{\backslash m}^{(i+1)})p(Att_m) \\ =\ & \lambda_m^{Att_m}(1-\lambda_m)^{1-Att_m} \prod_d \theta_{d|Att_m, \boldsymbol{att}_{\backslash m}^{(i+1)}}^{ext_d}(1-\theta_{d|Att_m, \boldsymbol{att}_{\backslash m}^{(i+1)}})^{1-ext_d}. \end{aligned}$$

Here, we used Bayes' theorem in the first line. The terms irrelevant to Att_m are eliminated in the second line. The remaining terms are expressed with their parameters in the third line. In the Appendix, we show a Gibbs sampling algorithm for the abstract argumentation model.

4. Correctness

This section discusses the relation between the abstract argumentation model and solutions to both of the inverse and forward problems of the abstract argumentation. We show that a solution to the inverse problem is equivalent to a maximum likelihood (ML) estimate, and that a solution to the forward problem is equivalent to an ML prediction. We next discuss the finding that the ML estimate and prediction are both special cases of probabilistic inference on the abstract argumentation model. Finally, we show experiments on the inverse problem, using a synthetic dataset.

4.1. Solutions to Inverse Problems

For simplicity, we do not distinguish attack relation $att \subseteq arg \times arg$ and value sequence $\boldsymbol{att}$ of attack-relation variables when $(Att_{(a,b)} = 1) \in \boldsymbol{att}$ iff $(a,b) \in att$ and $(Att_{(a,b)} = 0) \in \boldsymbol{att}$ iff $(a,b) \notin att$, for all $a, b \in arg$. Similarly, we do not distinguish extension set $S \subseteq Pow(arg)$ and value sequence $\boldsymbol{ext}$ of extension variables when $(Ext_s = 1)$ iff $s \in S$ and $(Ext_s = 0)$ iff $s \notin S$, for all $s \in Pow(arg)$.

Given extensions $\boldsymbol{ext}$, an attack relation $\hat{\boldsymbol{att}}$ is known as an ML estimate if it satisfies the following equation.

$$\hat{\boldsymbol{att}} = \arg\max_{\boldsymbol{att}} p(\boldsymbol{ext}|\boldsymbol{att})$$

Recall that a semantics ε, a set arg of arguments and a known attack relation $\widetilde{att}$ are all arbitrary but fixed. The following two theorems state that a solution to the inverse problem is an ML estimate in $\mathcal{M}$, but not vice versa.

Theorem 4.1. *Given extensions* $\boldsymbol{ext}$*, if attack relation* $\boldsymbol{att}$ *is a solution to the inverse problem then* $\boldsymbol{att}$ *is an ML estimate in* $\mathcal{M}$*.*

Proof. See the Appendix. □

The converse of Theorem 4.1 does not hold in general.

Proposition 4.2. *Given extensions* $\boldsymbol{ext}$*, if attack relation* $\boldsymbol{att}$ *is an ML estimate in* $\mathcal{M}$ *then* $\boldsymbol{att}$ *is not necessarily a solution to the inverse problem.*

Proof. It is enough to show a counterexample. Given $\boldsymbol{ext} = \{\emptyset, \{a, b\}\}$ and $\widetilde{\boldsymbol{att}} = \emptyset$, there is an ML estimate $att_{(a,b)}$ since $p(\boldsymbol{ext}|\widetilde{\boldsymbol{att}}, att_{(a,b)}) \geq 0$ holds for any $att_{(a,b)}$. However, it is never a solution to the inverse problem since there is no attack relation of which $\boldsymbol{ext}$ is the set of extensions. □

These two theorems imply that a solution to an ML estimate is weaker than that of an inverse problem. Here, 'weak' does not mean 'worthless'. The weakness of an ML estimate allows us to consider the presence of noise and the multiplicity of solutions.

Example 4.3. Let $arg = \{a, b\}$, $\boldsymbol{ext} = \{\emptyset, \{a, b\}\}$ and $\widetilde{att} = \emptyset$. Given $\boldsymbol{ext}$, there is no solution to the inverse problem as no attack relation yields $\boldsymbol{ext}$. However, $(Att_{(a,b)} = 1, Att_{(b,a)} = 1)$ is the ML estimate because we have

$$
\begin{aligned}
p(Ext_{\emptyset} = 1, Ext_{\{a,b\}} = 1 | Att_{(a,b)} = 0, Att_{(b,a)} = 0) &\propto \theta_{\emptyset|0,0}\theta_{\{a,b\}|0,0} = 0 \\
p(Ext_{\emptyset} = 1, Ext_{\{a,b\}} = 1 | Att_{(a,b)} = 1, Att_{(b,a)} = 0) &\propto \theta_{\emptyset|1,0}\theta_{\{a,b\}|1,0} = \frac{1}{9} \\
p(Ext_{\emptyset} = 1, Ext_{\{a,b\}} = 1 | Att_{(a,b)} = 0, Att_{(b,a)} = 1) &\propto \theta_{\emptyset|0,1}\theta_{\{a,b\}|0,1} = \frac{1}{9} \\
p(Ext_{\emptyset} = 1, Ext_{\{a,b\}} = 1 | Att_{(a,b)} = 1, Att_{(b,a)} = 1) &\propto \theta_{\emptyset|1,1}\theta_{\{a,b\}|1,1} = \frac{1}{3}.
\end{aligned}
$$

Here, we assumed the attack parameters $\lambda_{(a,b)} = \lambda_{(b,a)} = 1/2$ and exponential parameter $\theta_{d|\boldsymbol{att}}$ defined with $w = 2$ and complete semantics.

Example 4.3 shows that the ML estimate is more useful than the solution to the inverse problem. The reason is that it gives an answer regardless of the presence or absence of a noise in the observation.

We can benefit more from the model $\mathcal{M}$ due to the fact that an ML estimate is an approximation to the posterior distribution $p(\boldsymbol{Att}|\boldsymbol{ext})$. Since it is a probability distribution, it can be written as an N-tuple

$$
p(\boldsymbol{Att}|\boldsymbol{ext}) = \langle p(\boldsymbol{att}_1|\boldsymbol{ext}), p(\boldsymbol{att}_2|\boldsymbol{ext}), \cdots, p(\boldsymbol{att}_N|\boldsymbol{ext}) \rangle,
$$

where N is the number of possible different attack relations. Now, we assume that the posterior distribution has a sharp peak at an attack relation, denoted by $\hat{\boldsymbol{att}}$. It means that $\hat{\boldsymbol{att}}$ is very likely when given the extensions. Under the assumption, we have $p(\boldsymbol{Att}|\boldsymbol{ext}) \simeq 1$ if $\boldsymbol{Att} = \hat{\boldsymbol{att}}$ and $p(\boldsymbol{Att}|\boldsymbol{ext}) \simeq 0$ otherwise, where $\simeq$ denotes an

approximation. $\hat{\boldsymbol{att}}$ is now expressed as follows.

$$\begin{aligned}\hat{\boldsymbol{att}} &= \arg\max_{\boldsymbol{att}} p(\boldsymbol{att}|\boldsymbol{ext}) = \arg\max_{\boldsymbol{att}} \frac{p(\boldsymbol{ext}|\boldsymbol{att})p(\boldsymbol{att})}{p(\boldsymbol{ext})} \\ &= \arg\max_{\boldsymbol{att}} p(\boldsymbol{ext}|\boldsymbol{att})p(\boldsymbol{att})\end{aligned}$$

Here, $\hat{\boldsymbol{att}}$ is known as a maximum a posteriori (MAP) estimate, which maximises the posterior probability, i.e., $p(\hat{\boldsymbol{att}}|\boldsymbol{ext})$.

Next, we further assume that the prior distribution over attack relations is a uniform distribution. It means that we assume no information about the presence or absence of attack relations. Let c be a constant such that $p(\boldsymbol{att}) = c$, for all $\boldsymbol{att}$. We then have

$$\hat{\boldsymbol{att}} = \arg\max_{\boldsymbol{att}} p(\boldsymbol{ext}|\boldsymbol{att})c = \arg\max_{\boldsymbol{att}} p(\boldsymbol{ext}|\boldsymbol{att}).$$

Here, $\hat{\boldsymbol{att}}$ is known as an ML estimate, which maximises the likelihood function, i.e., $p(\boldsymbol{ext}|\boldsymbol{att})$.

In sum, we saw that a MAP estimate is an approximation to the posterior distribution in the sense that they are equivalent under the assumption that the posterior distribution over attack relations has a sharp peak. We also saw that an ML estimate is an approximation to a MAP estimate in the sense that they are equivalent under the assumption that the prior distribution over attack relations is a uniform distribution. The following practical implications follow from these facts.

- One should use an ML estimate instead of a non-probabilistic approach because it gives a solution regardless of the presence or absence of noises in observations.
- One should use a MAP estimate instead of an ML estimate because it allows us to consider a prior belief on the presence or absence of attacks between arguments.
- One should use the posterior distribution instead of a MAP estimate because it tells us the uncertainty about the extent to which each attack relation is likely to be the case.

4.2. Solutions to Forward Problems

We next investigate the relation between model $\mathcal{M}$ and solutions to the forward problem of the abstract argumentation. Given an attack relation $\boldsymbol{att}$, set $\hat{\boldsymbol{ext}}$ of extensions is known as an ML prediction if it satisfies the following equation.

$$\hat{\boldsymbol{ext}} = \arg\max_{\boldsymbol{ext}} p(\boldsymbol{ext}|\boldsymbol{att})$$

The following theorem states that an ML prediction in $\mathcal{M}$ is equivalent to a solution to the forward problem.

Theorem 4.4. *Given attack relation* $\boldsymbol{att}$*, set* $\boldsymbol{ext}$ *of extensions is a solution to the forward problem in* $\mathcal{M}$ *iff* $\boldsymbol{ext}$ *is an ML prediction with* $w \geq 2$*.*

Now, we ask how probabilistic inference on $\mathcal{M}$ extends solutions to the forward problem. To answer this, we derive the solution to the forward problem from the evidence (or marginal likelihood), which is the most general concept given by probabilistic

inference on $\mathcal{M}$ for the forward problem. The evidence is given by

$$p(\boldsymbol{Ext}) = \sum_{\boldsymbol{att}} p(\boldsymbol{Ext}, \boldsymbol{att}) = \sum_{\boldsymbol{att}} p(\boldsymbol{Ext}|\boldsymbol{att})p(\boldsymbol{att}).$$

Here, $\sum_{\boldsymbol{att}}$ is the abbreviation for $\sum_{att_{m_1}} \sum_{att_{m_2}} \cdots \sum_{att_{m_N}}$, for all $att_{m_n} \in \boldsymbol{att}$ where $1 \leq n \leq N$. Let us assume that the prior distribution over attack relations has a sharp peak, meaning that there is a very likely attack relation. Let $\hat{\boldsymbol{att}}$ denote the attack relation. Then, $p(\boldsymbol{Att}) \simeq 1$ if $\boldsymbol{Att} = \hat{\boldsymbol{att}}$ and $p(\boldsymbol{Att}) \simeq 0$ otherwise. We thus have

$$\sum_{\boldsymbol{att}} p(\boldsymbol{Ext}|\boldsymbol{att})p(\boldsymbol{att}) \simeq p(\boldsymbol{Ext}|\hat{\boldsymbol{att}}).$$

Here, $p(\boldsymbol{Ext}|\hat{\boldsymbol{att}})$ is known as a likelihood distribution.

Next, we further assume that the likelihood distribution has a sharp peak, meaning that there is a very likely set of extensions when given the attack relation. Then, $p(\boldsymbol{Ext}|\hat{\boldsymbol{att}}) \simeq 1$ if $\boldsymbol{Ext} = \hat{\boldsymbol{ext}}$ and $p(\boldsymbol{Ext}|\hat{\boldsymbol{att}}) \simeq 0$ otherwise. $\hat{\boldsymbol{ext}}$ is now expressed as an ML prediction as follows.

$$\hat{\boldsymbol{ext}} = \arg\max_{\boldsymbol{ext}} p(\boldsymbol{ext}|\hat{\boldsymbol{att}})$$

In sum, we saw that the likelihood distribution is an approximation to the evidence in the sense that they are equivalent under the assumption that the prior distribution over attack relations has a sharp peak. We then saw that an ML prediction is an approximation to the likelihood distribution in the sense that they are equivalent under the assumption that the likelihood distribution has a sharp peak. The following practical implications follow from these facts.

- There is no positive reason to use an ML prediction instead of a non-probabilistic approach to solve the forward problem.
- One should use the likelihood distribution instead of an ML prediction because it tells us the uncertainty about the extent to which each set of arguments is an extension.
- One should use the evidence instead of the likelihood distribution because it allows us to discuss extensions caused by more than one uncertain attack relation.

4.3. Experiments

The abstract argumentation model was discussed in terms of a theoretical point of view in Sections 4.1 and 4.2. In this section, we empirically discuss the correctness of its solutions to the inverse problems by using synthetic datasets. Each dataset is generated from the extensions of a randomly generated AF (abstract argumentation framework) with different assumptions. We add different amounts of noise to the extensions and then feed the noisy extensions to the abstract argumentation model. The model predicts the attack relation of the AF. We discuss the correctness of the prediction under different noise levels, acceptability semantics, numbers of arguments and restrictions on attack relations.

A noise is either subtractive or additive. A noise is subtractive if it causes a subtraction of random sets of arguments from the set of the extensions. A noise is additive

if it causes an addition of random sets of arguments to the set of extensions. As the constant w of the exponential parameter becomes larger, the likelihood distribution becomes sharper. It generally contributes to the convergence of the attack-relation distribution. However in practice, large w often causes division by zero in probabilistic inference. The occurrence of the error depends on the cardinality of the dataset, which is affected by the noise types and levels. We manually set a large enough w that does not cause the error, for all noise types and levels. We introduce no prior knowledge about the presence of an attack relation. We thus assumed the uniform prior distributions over attack relations. We assumed the number of iterations $I = 500$ and the burning period $B = 0$ in the approximate inference.

A true positive example is a positive example correctly classified as being positive whereas a true negative example is a negative example correctly classified as being negative. A false positive example is a negative example incorrectly classified as being positive whereas a false negative example is a positive example incorrectly classified as being negative. TP, TN, FP and FN denote the numbers of true positive examples, true negative examples, false positive examples and false negative examples, respectively. Accuracy, precision and recall are defined as follows.

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + FN + FP + TN} \\ Precision &= \frac{TP}{TP + FP} \\ Recall &= \frac{TP}{TP + FN} \end{aligned}$$

F-measure is defined as the harmonic mean of precision and recall.

$$F\text{-}measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

F-measure is more appropriate than accuracy, as the data (on the presence or absence of attacks) used to generate the dataset is skewed in our experiments. Figure 5 shows the average expected values or the average values of the F-measure of ten synthetic datasets. The top left figure allows us to compare different estimations available in the abstract argumentation model. For example, each F-measure value on the curve with the caption 'Bayes & Stable' was obtained by using the posterior attack-relation distribution estimated by using stable semantics. Each F-measure value on the curve with the caption 'MAP & Complete' was obtained by using the attack relation with a MAP estimated by using complete semantics. Each dataset was generated from an AF with an argument set with the cardinality of ten and a symmetric and irreflexive attack relation with the cardinality of ten. It is observed that the use of stable semantics still keeps the F-measure values of over 0.8 given the datasets with the noise of 40%. Complete semantics is inferior to stable semantics in the F-measure and there is a difference between the Bayesian and MAP estimations with complete semantics. The result suggests that a relatively large number of data on complete semantics makes the posterior distribution difficult to converge.

The top right figure in Figure 5 allows us to compare the F-measure values on different numbers of arguments. Given stable semantics, each F-measure value on the curve with the caption '10 args & Bayes' was obtained by using the posterior attack-relation distribution on ten arguments. It is thus equivalent to the curve with the

caption 'Bayes & Stable' shown on the top left figure in Figure 5. Each F-measure value on the curve with the caption '15 args & MAP' was obtained by using the attack relation with a MAP probability on fifteen arguments. Each dataset on ten (resp. fifteen) arguments was generated using stable semantics and an AF with an argument set with the cardinality of ten (resp. fifteen) and a symmetric and irreflexive attack relation with the cardinality of ten (resp. fifteen).

The bottom left figure in Figure 5 allows us to compare the effects of the subtractive noise and additive noise. As the noise becomes heavier, the F-measure values in the subtractive noise become worse than the ones in the additive noise. The reason is that the increase in the subtractive noise is equivalent to the approach to the empty dataset, which provides nothing with the model. Each dataset was generated using stable semantics and an AF with an argument set with the cardinality of fifteen and a symmetric and irreflexive attack relation with the cardinality of fifteen.

The bottom right figure in Figure 5 allows us to compare the F-measure values in the prediction of attack relations with and without symmetricity. Each dataset was generated using stable semantics, subtractive noises and an AF with an argument set with the cardinality of ten and an irreflexive attack relation with or without symmetricity, with the cardinality of ten. The result shows that the abstract argumentation model is totally useless for the prediction of an attack relation without symmetricity. The reason is that, as discussed in Proposition 2.5, there are generally multiple AFs resulting in the same set of extensions when the attack relations are not symmetric. In practice, this causes the divergence of the posterior distribution in the approximate inference.

4.4. Application Example

Figure 6 shows data, denoted by D, about the acceptability of arguments. We manually extracted the ten arguments from an online forum.[4] We then showed the ten arguments to twenty-nice individuals and collected their sentiments regarding the acceptability of each argument anonymously. The input to Algorithm 1 is as follows: $\boldsymbol{ext} = D$, $\varepsilon = c$, $w = 100$, $\lambda_{\{a,b\}} = 0.5$ for all arguments a and b, $I = 100$ and $B = 0$.

Figure 7 shows the all attack relations sampled during one hundred iterations of the Gibbs sampling procedure. Notably, the algorithm very frequently sampled the first three attack relations, which all agree the presence of attacks $\{a, h\}$, $\{b, d\}$, $\{b, g\}$ and $\{c, e\}$. They are all intuitively reasonable. For $\{a, h\}$, a states '*euthanasia (painless death) should be allowed by law because doctors should respect patient's will in medical treatment*', and h states '*patient's will for death is not enough to apply euthanasia.*' The attack $\{a, h\}$ would be more straightforward if we consider the following arguments. Indeed, g states '*it is possible that doctors apply euthanasia when their (i.e., patient's and her family's) will for euthanasia is confirmed.*' For $\{b, d\}$, b states '*I doubt doctor's right to commit a murder*', and d states '*only doctors can apply euthanasia appropriately.*' For $\{b, g\}$, b is the same as shown above and g states '*it is possible that doctors apply euthanasia.*' For $\{c, e\}$, c states '*... but disagree with the point that one who applies euthanasia is a doctor*', and e states '*it will be scary if there are professionals for euthanasia.*' Here, it is straightforward from the series of argumentation that 'professionals' means professionals other than doctors.

[4]See Appendix for the textual contents of the ten arguments.

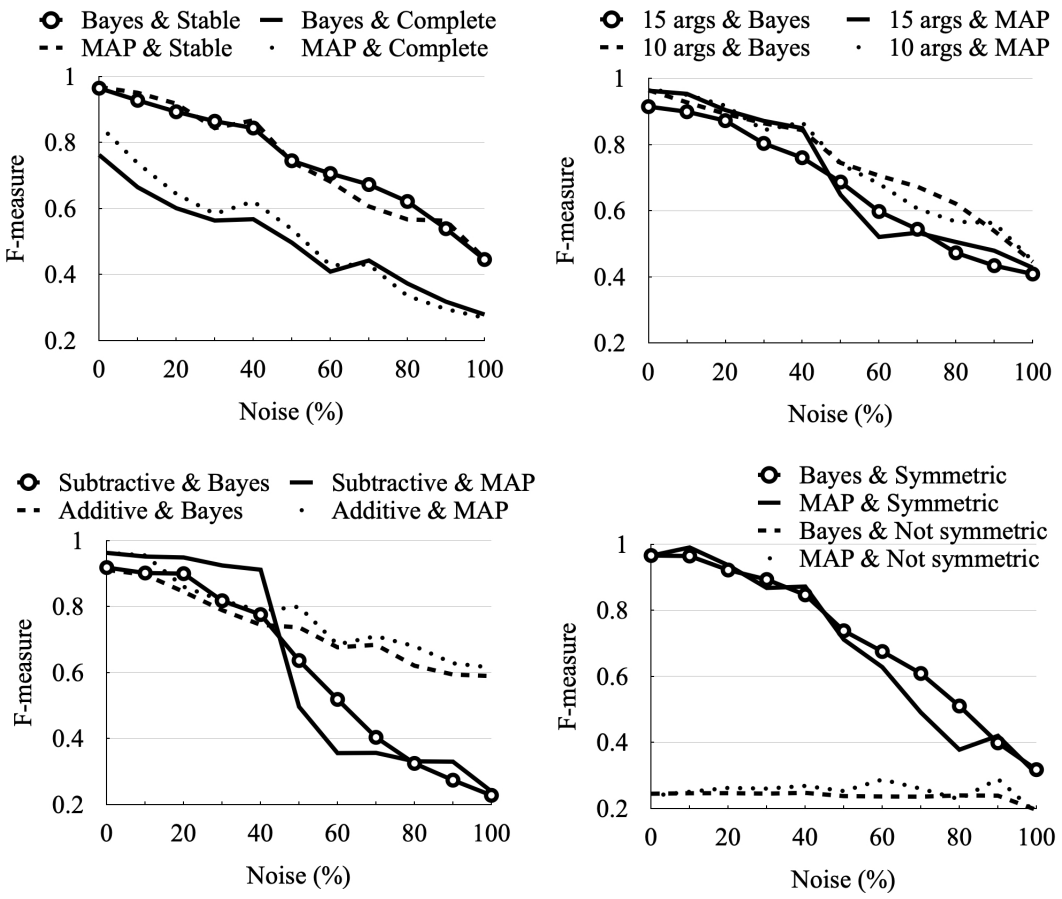


Figure 5. The average expected values or average values of the F-measure, which shows the correctness of predictions in the inverse problems. The horizontal axes are the amounts of noises in percentage.

5. Conclusions and Discussion

This paper provided an abstract argumentation model that gives broad views of solutions to the forward and inverse problems of the abstract argumentation. Given sets of acceptable arguments, the inverse problem is to find attack relations that explain the acceptability in terms of the acceptability semantics. It is the inverse of the forward problem that is the traditional problem; aiming to find the acceptability of arguments from an attack relation by using the acceptability semantics. We showed that solutions to the inverse and forward problems are equivalent to an ML estimation and ML prediction available in the model, respectively. They are special cases of the posterior distribution and the evidence, respectively, which are both obtained by probabilistic inference on the model.

From a statistical point of view, our abstract argumentation model is called a mixture model as it is a sequential combination of two types of distributions; one for attack relations and the other for extensions. Following the abstract argumentation, our model assumes that all extensions are generated from the same attack relation. It is different from the typical model, e.g., the mixtures of Bernoulli distributions and Gaussian distributions (Bishop, 2006). Indeed, it assumes that each element of the child distribution can be generated from a different element of the parent distribution.

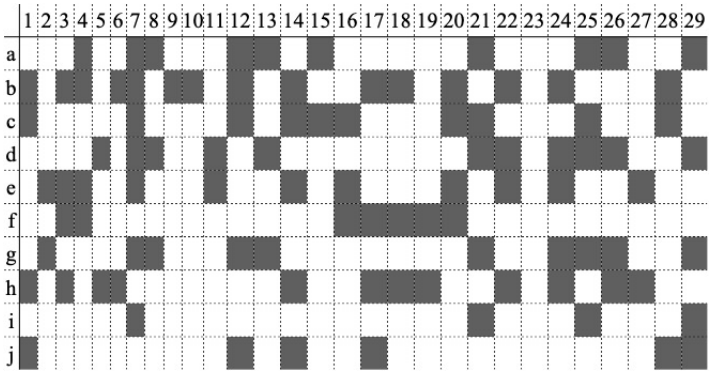


Figure 6. Twenty-nine anonymous participants' sentiments regarding acceptability of ten individual arguments on active euthanasia. Each white and grey cell denotes that the participant agrees and disagrees the argument respectively.

Besides, a random variable is assigned to each pair of arguments in the abstract argumentation model. We could have assigned a random variable to a directed graph, i.e., an attack relation on the set of arguments. Such a model and the abstract argumentation model behave in exactly the same way in the exact probabilistic inference. However, we decided to use the latter as a Gibbs sampling method is not available in the former.

The past two decades in the field of computational argumentation in AI have witnessed an intensive study of acceptability semantics and dialectical proof theories. They give various interpretations and derivations of the acceptability of arguments when argumentative knowledge is represented as argumentation frameworks, e.g., (Amgoud, 2009; Baroni et al., 2005; Bench-Capon, 2002; Caminada, 2006; Cayrol and Lagasquie-Schiex, 2005; Coste-Marquis et al., 2005; Dung et al., 2006; Leite and Martins, 2011; Modgil and Luck, 2009; Verheij, 1996). On the other hand, another emerging research direction is to find, complement or revise an argumentation framework from the acceptability of arguments, e.g., (Kido, 2018; Kido and Okamoto, 2017; Niskanen et al., 2016, 2018, 2019; Riveret and Governatori, 2016). In particular, the authors (Kido, 2018; Kido and Okamoto, 2017) introduce a generative model of acceptability of arguments into computational argumentation and use it to solve an inverse problem of the abstract argumentation. The generative model formalises the inferential process by which acceptability of arguments is probabilistically generated from attack relations by following the acceptability semantics. The term 'inverse problem' the authors introduced in (Kido and Liao, 2019) is gaining popularity, e.g., see (Nir Oren et al., 2022). However, since the generative model lacks generality, it is still not clear what kinds of general problems should be solved and what kinds of problems can be solved based on the idea of generative models. In Section 2, we thus defined the forward and inverse problems of the abstract argumentation as the problems to be solved and discussed in the presence of noise. In Section 3, we then gave a simple but general generative model for the solutions to the problems. In Sections 4.1 and 4.2, we statistically characterised the properties of the solutions provided by probabilistic inference on the generative model. We finally demonstrated its applicability in argumentation mining in Sections 4.3 and 4.4.

To the best of our knowledge, this paper for the first time offers a general discussion about a generative model of acceptability of arguments in the context of computational

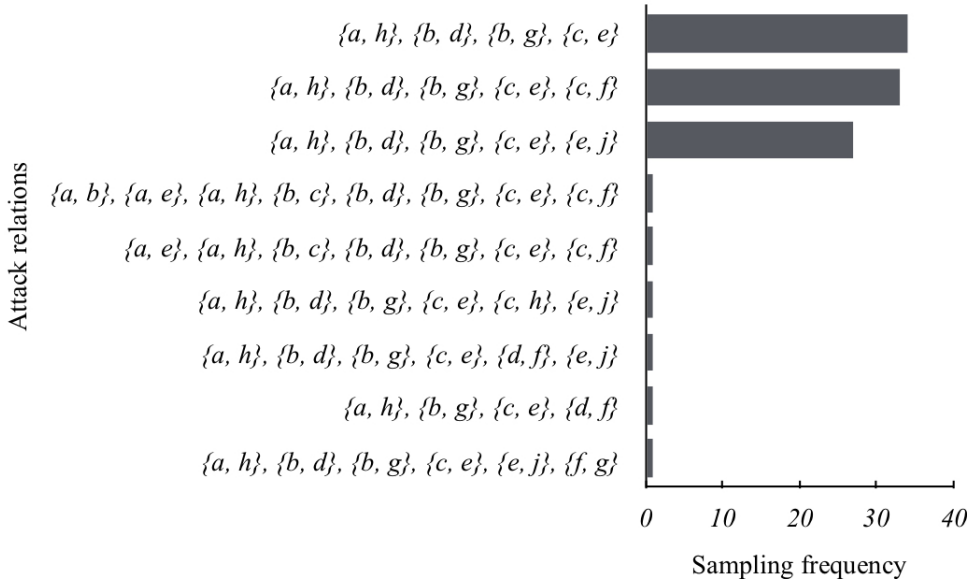


Figure 7. Sampled attack relations versus the number of samples.

argumentation. The inverse problem cannot be solved properly without the theory of the forward problem. This paper thus also contributes to finding another value of these formalisms of the forward problem for their inverse problems. From a data science point of view, a weakness of the study of the abstract argumentation, and of symbolic AI in general, is the knowledge acquisition bottleneck on how to acquire knowledge from data. This is because, in general, a knowledge representation is not what one assumes, but what one wants and can obtain as a result of problem analysis. The input to the inverse problem, in practice, is a set of sentiments regarding the acceptability of arguments. In contrast to an attack relation required in the forward problem, it is unstructured data and thus it can be collected from the web more easily, e.g., via votes in various social networking services.

At the same time, the most important future work for practical applications is the scalability of the abstract argumentation model. We implemented the Gibbs sampling algorithm for approximate probabilistic inference on the abstract argumentation model. However, our model still cannot handle hundreds or thousands of arguments in the inverse problem. The most promising approach would be a stochastic block model, which is a statistical approach to graph and digraph clustering. The key idea is that the probability distribution over an edge between two nodes is assumed to depend only on the probability distribution over an edge between the latent groups to which they belong. Here, the number of latent groups is assumed to be much smaller than the number of nodes. It is worthwhile investigating whether the idea is effective in the inverse abstract argumentation problem. The abstract argumentation model serves as a basis for developing the stochastic block model of the abstract argumentation.

Acknowledgement

The authors are grateful to Martin Caminada for valuable discussion. The authors report there are no competing interests to declare.

References

Amgoud, L. (2009). Repairing preference-based argumentation frameworks. In *Proc. of the 21st International Joint Conference on Artificial Intelligence*, pages 665–670.

Aster, R. C., Borchers, B., and Thurber, C. (2004). *Parameter Estimation and Inverse Problem*. Academic Press.

Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A., and Slonim, N. (2017a). Stance classification of context-dependent claims. In *Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Volume 1*, pages 251–261.

Bar-Haim, R., Edelstein, L., Jochim, C., and Slonim, N. (2017b). Improving claim stance classification with lexical knowledge expansion and context utilization. In *Proc. of the 4th Workshop on Argument Mining*, pages 32–38.

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

Bench-Capon, T. J. M. (2002). Value-based argumentation frameworks. In *Proc. of the 9th International Workshop on Non-Monotonic Reasoning*, pages 443–454.

Bench-Capon, T. J. M. and Dunne, P. E. (2007). Argumentation in artificial intelligence. *Artificial Intelligence*, 171(10–15):619–641.

Bex, F. and Renooij, S. (2016). From arguments to constraints on a bayesian network. computational models of argument. In *Proc. of the 6th International Conference on Computational Models of Argument*, pages 95–106.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Boltužić, F. and Šnajder, J. (2014). Back up your stance: Recognizing arguments in online discussions. In *Proc. of the 1st Workshop on Argumentation Mining*, pages 49–58.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *arXiv:1508.05326 [cs.CL]*.

Cabrio, E. and Villata, S. (2013). A natural language bipolar argumentation approach to support users in online debate interactions. *Argumentation & Computation*, 4(3):209–230.

Caminada, M. (2006). Semi-stable semantics. In *Proc. of the 1st International Conference on Computational Models of Argument*, pages 121–130.

Cayrol, C. and Lagasquie-Schiex, M. C. (2005). On the acceptability of arguments in bipolar argumentation frameworks. In *Proc. of the 8th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 378–389.

Coste-Marquis, S., Devred, C., and Marquis, P. (2005). Prudent semantics for argumentation frameworks. In *Proc. of the 17th International Conference on Tools with Artificial Intelligence*, pages 568–572.

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

Dung, P. M., Mancarella, P., and Toni, F. (2006). A dialectic procedure for sceptical, assumption-based argumentation. In *Proc. of the 1st International Conference on Computational Models of Argument*, pages 145–156.

Dunne, P. E., Dvořák, W., Linsbichler, T., and Woltran, S. (2015). Characteristics of multiple viewpoints in abstract argumentation. *Artificial Intelligence*, 228:153–178.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(1):721–741.

Grabmair, M., Gordon, T. F., and Walton, D. (2010). Probabilistic semantics for the carneades

argument model using bayesian networks. In *Proc. of the 3rd International Conference on Computational Models of Argument*, pages 255–266.

Hunter, A. (2020). Generating instantiated argument graphs from probabilistic information. In *Proc. of the 24th European Conference on Artificial Intelligence*, pages 769–776.

Kido, H. (2018). Bayesian model selection in statistical construction of justification. In *Proc. of the 16th International Conference on Knowledge Representation and Reasoning*, pages 647–648.

Kido, H. and Liao, B. (2019). A bayesian approach to direct and inverse abstract argumentation problems. In *arXiv:1909.04319 [cs.AI]*.

Kido, H. and Okamoto, K. (2017). A Bayesian approach to argument-based reasoning for attack estimation. In *Proc. of the 26th International Joint Conference on Artificial Intelligence*, pages 249–255.

Lawrence, J. and Reed, C. (2016). Argument mining using argumentation scheme structures. In *Proc. of the 6th International Conference on Computational Models of Argument*, pages 379–390.

Leite, J. and Martins, J. (2011). Social abstract argumentation. In *Proc. of the 22nd international joint conference on Artificial Intelligence*, pages 2287–2292.

Levesque, H. J., Davis, E., and Morgenstern, L. (2012). The winograd schema challenge. In *Proc. of the 13th international Conference on Principles ofKnowledge Representation and Reasoning*, pages 552–561.

Lippi, M. and Torroni, P. (2015). Context-independent claim detection for argument mining. In *Proc. of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 185–191.

Lippi, M. and Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2):10:1–10:25.

MacCartney, B. and Manning, C. D. (2007). Natural logic for textual inference. In *Proc. of the Workshop on Textual Entailment and Paraphrasing*, pages 193–200.

Mayer, T., Cabrio, E., Lippi, M., Torroni, P., and Villata, S. (2018). In *Proc. of the 7th International Conference on Computational Models of Argument*, pages 137–148.

Modgil, S. and Luck, M. (2009). Argumentation based resolution of conflicts between desires and normative goals. In *Proc. of the 5th International Workshop on Argumentation in Multi-Agent Systems*, pages 19–36.

Nielsen, S. H. and Parsons, S. (2007). An application of formal argumentation: Fusing bayesian networks in multi-agent systems. *Artificial Intelligence*, 171:754–775.

Nir Oren, B. Y., Vesic, S., and Baptista, M. (2022). Inverse problems for gradual semantics. In *Proc. of the 31st International Joint Conference on Artificial Intelligence*, pages 2719–2725.

Niskanen, A., Wallner, J. P., and Järvisalo, M. (2016). Synthesizing argumentation frameworks from examples. In *Proc. of the 22nd European Conference on Artificial Intelligence*, pages 551–559.

Niskanen, A., Wallner, J. P., and Järvisalo, M. (2018). Extension enforcement under grounded semantics in abstract argumentation. In *Proc. of the 16th International Conference on Knowledge Representation and Reasoning*, pages 178–183.

Niskanen, A., Wallner, J. P., and Järvisalo, M. (2019). Synthesizing argumentation frameworks from examples. *Journal of Artificial Intelligence Research*, 66:503–554.

Palau, R. M. and Moens, M. F. (2009). Argumentation mining: The detection, classification and structure of arguments in text. In *Proc. of the 12th International Conference on Artificial Intelligence and Law*, pages 98–107.

Prakken, H. and Vreeswijk, G. (2001). *Logics for Defeasible Argumentation*, volume 4, pages 219–318. Springer, handbook of philosophical logic, 2nd edition.

Riveret, R. (2020). On searching explanatory argumentation graphs. *Journal of Applied Non-Classical Logics*, 30(2):123–192.

Riveret, R. and Governatori, G. (2016). On learning attacks in probabilistic abstract argumentation. In *Proc. of the 15th International Conference on Autonomous Agents and Multiagent Systems*, pages 653–661.

Saha, S. and Sen, S. (2004). A bayes net approach to argumentation. In *Proc. of the 19th national conference on Artificial intelligence*, pages 966–967.

Saint-Dizier, P. (2018). A knowledge-based approach to warrant induction. In *Proc. of the 7th International Conference on Computational Models of Argument*, pages 289–300.

Silva, V. S., Freitas, A., and Handschuh, S. (2018). Recognizing and justifying text entailment through distributional navigation on definition graphs. In *Proc. of the 22nd AAAI Conference on Artificial Intelligence*.

SYNCLON (2013). Synclon$^3_\beta$. http://synclon3.com/, Retrieved May 2015.

Timmer, S. T., Meyer, J.-J. C., Prakken, H., Renooij, S., and Verheij, B. (2015). Explaining bayesian networks using argumentation. In *Proc. of the 13th European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 83–92.

Toledo-Ronen, O., Bar-Haim, R., and Slonim, N. (2016). Expert stance graphs for computational argumentation. In *Proc. of the 3rd Workshop on Argument Mining*, pages 119–123.

Uspensky, J. (1937). *Introduction to Mathematical Probability*. NcGraw-Hill Inc., US.

Verheij, B. (1996). Two approaches to dialectical argumentation: admissible sets and argumentation stages. In *Proc. of the 8th Dutch Conference on Artificial Intelligence*, pages 357–368.

Vreeswijk, G. A. (2005). Argumentation in bayesian belief networks. In *Proc. of the 2nd International Workshop on Argumentation in Multi-Agent Systems*, pages 111–129.

Zhao, K., Huang, L., and Ma, M. (2017). Textual entailment with structured attentions and composition. In *arXiv:1701.01126 [cs.CL]*.

Appendix A. Proofs of Theorems

Proposition 2.5. This directly follows from the fact that, for any two abstract argumentation frameworks AF_1 and AF_2 with the same set of arguments, and symmetric and irreflexive attack relations, if $\varepsilon(AF_1) = \varepsilon(AF_2)$ then $AF_1 = AF_2$ for any semantics except grounded semantics. As for grounded semantics, it is obvious because the empty set is the grounded extension of any AF_i $(i = 1, 2)$ with a symmetric non-empty attack relation. For the remaining semantics, since AF_i is symmetric and irreflexive, it can be regarded as an undirected graph without self-loop where a node and an edge represent an argument and an attack between arguments, respectively. For all independent sets S of this graph, S is an admissible set of AF_i. There is thus a set $T \supseteq S$ such that T is a preferred extension AF_i. Since AF_i is symmetric, a set of arguments is a preferred extension if and only if it is a stable extension. It is moreover obvious that a preferred extension is a complete extension. If $AF_1 \neq AF_2$ then the set of independent sets of AF_1 does not coincide with that of AF_2. Therefore, the set of preferred extensions of AF_1 does not coincide with that of AF_2. □

Proposition 3.6. Let f be the normalized exponential function. We then have

$$\lim_{w\to\infty} f(x) = \lim_{w\to\infty} \frac{w^{x-|arg|} - \frac{1}{w^{|arg|}}}{1 - \frac{1}{w^{|arg|}}} = \lim_{w\to\infty} w^{x-|arg|} = \begin{cases} 1 & x = |arg| \\ 0 & otherwise. \end{cases}$$

□

Proposition 3.7. Let f be the normalized exponential function. Using l'Hôpital's

rule, we have

$$\lim_{w\to\infty} f(x) = \lim_{w\to 1} \frac{w^x - 1}{w^{|arg|} - 1} = \lim_{w\to 1} \frac{xw^{x-1}}{|arg|w^{|arg|-1}} = \frac{x}{|arg|}.$$

□

Theorem 4.1. Let $\widetilde{\boldsymbol{att}}$ be an arbitrary attack relation. $\boldsymbol{ext}$ can be divided into four disjoint sets.

$$\begin{aligned}
\boldsymbol{ext}_{11} &= \{(Ext_d = 1) \in \boldsymbol{ext} | d \in \varepsilon(arg, \widetilde{\boldsymbol{att}} \cup \boldsymbol{att})\} \\
\boldsymbol{ext}_{10} &= \{(Ext_d = 1) \in \boldsymbol{ext} | d \notin \varepsilon(arg, \widetilde{\boldsymbol{att}} \cup \boldsymbol{att})\} \\
\boldsymbol{ext}_{01} &= \{(Ext_d = 0) \in \boldsymbol{ext} | d \in \varepsilon(arg, \widetilde{\boldsymbol{att}} \cup \boldsymbol{att})\} \\
\boldsymbol{ext}_{00} &= \{(Ext_d = 0) \in \boldsymbol{ext} | d \notin \varepsilon(arg, \widetilde{\boldsymbol{att}} \cup \boldsymbol{att})\}
\end{aligned}$$

Since acceptability is independent and identically distributed, we have

$$\begin{aligned}
& p(\boldsymbol{ext} | \widetilde{\boldsymbol{att}} \cup \boldsymbol{att}) \\
= \; & p(\boldsymbol{ext}_{11} | \widetilde{\boldsymbol{att}} \cup \boldsymbol{att}) p(\boldsymbol{ext}_{10} | \widetilde{\boldsymbol{att}} \cup \boldsymbol{att}) p(\boldsymbol{ext}_{01} | \widetilde{\boldsymbol{att}} \cup \boldsymbol{att}) p(\boldsymbol{ext}_{00} | \widetilde{\boldsymbol{att}} \cup \boldsymbol{att}) \\
= \; & \prod_{ext_{d1} \in \boldsymbol{ext}_{11}} \theta_{d1|\widetilde{\boldsymbol{att}} \cup \boldsymbol{att}} \prod_{ext_{d2} \in \boldsymbol{ext}_{10}} \theta_{d2|\widetilde{\boldsymbol{att}} \cup \boldsymbol{att}} \prod_{ext_{d3} \in \boldsymbol{ext}_{01}} (1 - \theta_{d3|\widetilde{\boldsymbol{att}} \cup \boldsymbol{att}}) \\
& \prod_{ext_{d4} \in \boldsymbol{ext}_{00}} (1 - \theta_{d4|\widetilde{\boldsymbol{att}} \cup \boldsymbol{att}}).
\end{aligned}$$

Since $\boldsymbol{att}$ is a solution to the inverse problem, i.e., $\boldsymbol{ext} = \varepsilon(arg, \widetilde{\boldsymbol{att}} \cup \boldsymbol{att})$, $\boldsymbol{ext}_{11} \cup \boldsymbol{ext}_{00} = \boldsymbol{ext}$ and $\boldsymbol{ext}_{01} \cup \boldsymbol{ext}_{10} = \emptyset$ holds. We thus have

$$p(\boldsymbol{ext} | \widetilde{\boldsymbol{att}} \cup \boldsymbol{att}) = \prod_{ext_{d1} \in \boldsymbol{ext}_{11}} \theta_{d1|\widetilde{\boldsymbol{att}} \cup \boldsymbol{att}} \prod_{ext_{d4} \in \boldsymbol{ext}_{00}} (1 - \theta_{d4|\widetilde{\boldsymbol{att}} \cup \boldsymbol{att}}).$$

Now, any attack relation that is not a solution to the inverse problem causes a shift of an element from $\boldsymbol{ext}_{11}$ to $\boldsymbol{ext}_{10}$ or $\boldsymbol{ext}_{00}$ to $\boldsymbol{ext}_{01}$. However, this never makes the probability higher. This is because $\theta_{d1|\widetilde{\boldsymbol{att}} \cup \boldsymbol{att}} > \theta_{d2|\widetilde{\boldsymbol{att}} \cup \boldsymbol{att}}$ and $(1 - \theta_{d4|\widetilde{\boldsymbol{att}} \cup \boldsymbol{att}}) > (1 - \theta_{d3|\widetilde{\boldsymbol{att}} \cup \boldsymbol{att}})$ hold from Definition 3.5 where $\theta_{d1|\widetilde{\boldsymbol{att}} \cup \boldsymbol{att}} = 1$, $\theta_{d2|\widetilde{\boldsymbol{att}} \cup \boldsymbol{att}} < 1$, $\theta_{d3|\widetilde{\boldsymbol{att}} \cup \boldsymbol{att}} = 1$ and $\theta_{d4|\widetilde{\boldsymbol{att}} \cup \boldsymbol{att}} < 1$ hold. □

Theorem 4.4. ($\Rightarrow$) $\boldsymbol{ext}$ can be divided into four disjoint sets.

$$\begin{aligned}
\boldsymbol{ext}_{11} &= \{(Ext_d = 1) \in \boldsymbol{ext} | d \in \varepsilon(arg, att)\} \\
\boldsymbol{ext}_{10} &= \{(Ext_d = 1) \in \boldsymbol{ext} | d \notin \varepsilon(arg, att)\} \\
\boldsymbol{ext}_{01} &= \{(Ext_d = 0) \in \boldsymbol{ext} | d \in \varepsilon(arg, att)\} \\
\boldsymbol{ext}_{00} &= \{(Ext_d = 0) \in \boldsymbol{ext} | d \notin \varepsilon(arg, att)\}
\end{aligned}$$

Since acceptability is independent and identically distributed, we have

$$\begin{aligned} p(\boldsymbol{ext}|\boldsymbol{att}) &= p(\boldsymbol{ext}_{11}|\boldsymbol{att})p(\boldsymbol{ext}_{10}|\boldsymbol{att})p(\boldsymbol{ext}_{01}|\boldsymbol{att})p(\boldsymbol{ext}_{00}|\boldsymbol{att}) \\ &= \prod_{ext_{d1}\in\boldsymbol{ext}_{11}} \theta_{d1|\boldsymbol{att}} \prod_{ext_{d2}\in\boldsymbol{ext}_{10}} \theta_{d2|\boldsymbol{att}} \prod_{ext_{d3}\in\boldsymbol{ext}_{01}} (1-\theta_{d3|\boldsymbol{att}}) \\ &\quad \prod_{ext_{d4}\in\boldsymbol{ext}_{00}} (1-\theta_{d4|\boldsymbol{att}}). \end{aligned}$$

Since $\boldsymbol{att}$ is a solution to the forward problem, $\boldsymbol{ext} = \varepsilon(arg, \boldsymbol{att})$ holds. Thus, $\boldsymbol{ext}_{11} \cup \boldsymbol{ext}_{00} = \boldsymbol{ext}$ and $\boldsymbol{ext}_{01} \cup \boldsymbol{ext}_{10} = \emptyset$ holds. We thus have

$$p(\boldsymbol{ext}|\boldsymbol{att}) = \prod_{ext_{d1}\in\boldsymbol{ext}_{11}} \theta_{d1|\boldsymbol{att}} \prod_{ext_{d4}\in\boldsymbol{ext}_{00}} (1-\theta_{d4|\boldsymbol{att}}) \quad \text{(A1)}$$

Now, any acceptability that is not a solution to the forward problem causes a shift of an element from $\boldsymbol{ext}_{11}$ to $\boldsymbol{ext}_{01}$ or $\boldsymbol{ext}_{00}$ to $\boldsymbol{ext}_{10}$. However, this never makes the probability higher because $\theta_{d1|\boldsymbol{att}} > (1-\theta_{d3|\boldsymbol{att}})$ and $(1-\theta_{d4|\boldsymbol{att}}) > \theta_{d2|\boldsymbol{att}}$ hold. This is because, from Definition 3.5, if $w \geq 2$ holds then $\theta_{d_1|\boldsymbol{att}} = 1$, $\theta_{d2|\boldsymbol{att}} < 0.5$, $\theta_{d3|\boldsymbol{att}} = 1$ and $\theta_{d4|\boldsymbol{att}} < 0.5$ hold. Here, we prove $\theta_{d|\boldsymbol{att}} < 0.5$ holds if $d \notin \varepsilon(arg, \boldsymbol{att})$ and $w \geq 2$, as follows. Let d be a set of arguments such that there is, at best, an extension $e \in \varepsilon(arg, att)$ satisfying $|sim(d, e)| = |arg| - 1$. Here, we do not need to think of $|sim(d, e)| < |arg| - 1$ because of the monotonicity of $\theta_{d|\boldsymbol{att}}$. We then have

$$\begin{aligned} \theta_{d|\boldsymbol{att}} &= \frac{w^{|arg|-1}-1}{w^{|arg|}-1} = \frac{1}{w}\frac{w^{|arg|}-w}{w^{|arg|}-1} \\ 2\theta_{d|\boldsymbol{att}} &= \frac{2}{w}\frac{w^{|arg|}-w}{w^{|arg|}-1}. \end{aligned}$$

Now, $2\theta_{d|\boldsymbol{att}} < 1$ holds, for all $w \geq 2$. Indeed, if $w = 2$ then $2/w = 1$ and $(w^{|arg|} - w)/(w^{|arg|} - 1) < 1$ hold. If $w > 2$ then $2/w < 1$ and $(w^{|arg|} - w)/(w^{|arg|} - 1) < 1$ hold as well.

($\Leftarrow$) We show that if $\boldsymbol{ext}$ is not a solution to the forward problem then it does not maximize the likelihood of $\boldsymbol{ext}$. We do not need to consider the case where there is no solution to a forward problem because it does not satisfy the antecedent. If $\boldsymbol{ext}$ is not a solution then $\boldsymbol{ext}_{11} \cup \boldsymbol{ext}_{00} \subset \boldsymbol{ext}$ and $\boldsymbol{ext}_{10} \cup \boldsymbol{ext}_{01} \supset \emptyset$ hold. However, since a forward problem satisfies the solution uniqueness, there is a unique $\boldsymbol{ext'}$ such that $\boldsymbol{ext'}_{11} \cup \boldsymbol{ext'}_{00} = \boldsymbol{ext'}$ and $\boldsymbol{ext'}_{10} \cup \boldsymbol{ext'}_{01} = \emptyset$. Since $\theta_{d1|\boldsymbol{att}} > (1-\theta_{d3|\boldsymbol{att}})$ and $(1-\theta_{d4|\boldsymbol{att}}) > \theta_{d2|\boldsymbol{att}}$ hold in Equation (A1), $p(\boldsymbol{ext'}|\boldsymbol{att}) > p(\boldsymbol{ext}|\boldsymbol{att})$ holds. □

Appendix B. Gibbs Sampling Algorithm

Algorithm 1 shows the Gibbs sampling algorithm for the abstract argumentation model. Lines 3-10 show how a value of every attack relation is generated based on the above formula. The algorithm iterates this process I times. In line 12-16, it constructs a histogram of the attack relations sampled after B iterations. It finally returns the normalized distribution of the histogram.

Algorithm 1 Gibbs sampling for the abstract argumentation model

Require: Observation $\boldsymbol{ext}$, semantics ε, constant w of the exponential parameters, constant λ_m of the attack parameters, iteration number I and burn-in period B
Ensure: Approximation $\hat{p}(\boldsymbol{Att}|\boldsymbol{ext})$ of the posterior distribution $p(\boldsymbol{Att}|\boldsymbol{ext})$

Get $\boldsymbol{att}^{(0)}$ by randomly assigning 0 or 1 to all elements of $\boldsymbol{Att}$
for $i = 0$ to I **do**
  for all $Att_m \in \boldsymbol{Att}$ **do**
    $\boldsymbol{prob} \leftarrow [1 - \lambda_m, \lambda_m]$ ▷ Compute $p(att_m^{(i+1)})$
    for all $ext_d \in \boldsymbol{ext}$ **do** ▷ Compute $p(ext_d|att_m^{(i+1)}, \boldsymbol{att}_{\setminus m}^{(i+1)})$
      $prob[0] \leftarrow prob[0] \cdot \theta_{d|Att_m=0,\boldsymbol{att}_{\setminus m}^{(i+1)}}^{ext_d}(1 - \theta_{d|Att_m=0,\boldsymbol{att}_{\setminus m}^{(i+1)}})^{1-ext_d}$
      $prob[1] \leftarrow prob[1] \cdot \theta_{d|Att_m=1,\boldsymbol{att}_{\setminus m}^{(i+1)}}^{ext_d}(1 - \theta_{d|Att_m=1,\boldsymbol{att}_{\setminus m}^{(i+1)}})^{1-ext_d}$
    end for
    $att_m^{(i+1)} \sim \boldsymbol{prob}$ ▷ Generate $att_m^{(i+1)}$ from $p(Att_m|\boldsymbol{att}_{\setminus m}^{(i+1)}, \boldsymbol{ext})$
  end for
end for
$\boldsymbol{freq} \leftarrow \emptyset$
for all $\boldsymbol{att} \in \{\boldsymbol{att}^{(i)}|B < i \leq I\}$ **do** ▷ Compute an attack relation histogram
  $count \leftarrow$ the number of occurrence of $\boldsymbol{att}$ in $(\boldsymbol{att}^{(i)}|B < i \leq I)$
  $\boldsymbol{freq} \leftarrow \boldsymbol{freq} \cup \{(count, \boldsymbol{att})\}$
end for
return $(count/(I - B)|(count, \boldsymbol{att}) \in \boldsymbol{freq})$

Appendix C. Argument Data

The ten arguments used in our empirical analysis have the following textual contents. They were presented in this order in (SYNCLON, 2013). We manually extracted and translated them into English.

a: *Laws should not allow euthanasia (painless death) because doctors should respect patient's will in medical treatment.*

b: *Laws should not allow euthanasia. You assume that one who applies euthanasia is a doctor. I doubt doctor's right to commit a murder.*

c: *I agree with you that euthanasia should be allowed by law, but disagree with the point that one who applies is a doctor. Doctors should always consider the way to cure diseases.*

d: *If doctor's role is only to cure diseases then they can do nothing for patients with an untreatable disease. I think that medical treatment should consider death more seriously. Only doctors can apply euthanasia appropriately because they can judge patients' physical and mental state accurately.*

e: *But, it will be scary if there are professionals for euthanasia.*

f: *I mean that I disagree with the point that doctors encourages a patient to choose euthanasia. Doctors can help patients with an untreatable disease without encouraging them to choose euthanasia. Doctors and patients can lay heads together to think about how they can live with a disease. This is how doctors can consider patients' death.*

g: *Of course, no one has a right to encourage patients to choose euthanasia. Euthanasia can be applied on the basis of the agreement by the patient and her family. I think it is possible that doctors apply euthanasia when their will for euthanasia is confirmed.*

h: *I think that the agreement or confirmation does not show patient's true will. I*

occasionally wish for death when I have a hard experience. But, it sometimes comes from a temporary emotion. I am sure that my experience is something little compared to patient's sufferance. But, I think that patient's will for death is not enough to apply euthanasia.

i: *Can you accept legal euthanasia if it is based not only on patient's will, but also her family's will?*

j: *I cannot accept legal euthanasia even though it is based not only on patient's will, but also her family's will. I cannot accept euthanasia without considering patient's physical condition. So, I agree with passive euthanasia.*