A Bayesian Approach to Argument-Based Reasoning for Attack Estimation

Hiroyuki Kido
Institute of Logic and Cognition
Sun Yat-sen University
kido@mail.sysu.edu.cn

Keishi Okamoto
Department of Information Systems
National Institute of Technology, Sendai College
okamoto@sendai-nct.ac.jp

Abstract

The web is a source of a large amount of arguments and their acceptability statuses (e.g., votes for and against the arguments). However, relations existing between the fore-mentioned arguments are typically not available. This study investigates the utilisation of acceptability semantics to statistically estimate an attack relation between arguments wherein the acceptability statuses of arguments are provided. A Bayesian network model of argument-based reasoning is defined in which Dung's theory of abstract argumentation gives the substance of Bayesian inference. The model correctness is demonstrated by analysing properties of estimated attack relations and illustrating its applicability to online forums.

1 Introduction

The internet is full of debates in which several individuals argue on various issues from different standpoints. Several arguments and their metadata (e.g., by whom and when as well as what arguments and votes are put forward) are increasingly available and reusable online. However, it is difficult for computers and even for individuals to identify relations, such as disputes and support and clarification relations, which exist between arguments because the existence of these types of relations is almost context-dependent and knowledge-dependent. Specifically, the detection of an attack relation is important because it can potentially impact sentiment analysis (or opinion mining) and persuasive technology. The primary aim of sentiment analysis involves providing an answer as to the type of opinions that are held by individuals. In contrast, the detection of an attack relation makes it possible to proceed further by providing reasons for why individuals hold opinions as well as how opinions can be changed.

Detection of an attack relation is one of the main challenges tackled by argumentation (or argument) mining. According to Moens [Moens, 2013], "Argumentation mining can be defined as the detection of the argumentative discourse structure in text or speech and the recognition or functional classification of the components of the argumentation". There are at least two approaches to argumentation mining. The first approach uses computational linguistics (or natural language processing) and machine learning. Given textual discourse, the goal involves identifying individual arguments, their internal structures, and their interactions [Palau and Moens, 2009; Lawrence and Reed, 2016]. The second approach uses computational argumentation and machine learning. Given acceptability statuses of individual arguments, the goal involves identifying an attack relation between the arguments such as the AF synthesis problem [Niskanen *et al.*, 2016] by considering realisability [Dunne *et al.*, 2015] and the abstract structure learning [Riveret and Governatori, 2016] by considering probability theory.

This study corresponds to the second approach. The last two decades in computational argumentation witnessed intensive studies that used acceptability semantics to define acceptability statuses of individual arguments given an attack relation between those arguments. By contrast, the present study asks the following question: Given acceptability statuses of individual arguments, how should acceptability semantics be utilised to identify an attack relation between the fore-mentioned arguments? This question is practically interesting because these types of statuses are available online such as "Vote" in online forums, "Like" in Facebook, "Helpful" in Amazon customer reviews, and "Useful and clear" in Stack Overflow.

In the study, a Bayesian network model of argument-based reasoning is defined. It provides argument-based Bayesian inference in which Dung's acceptability semantics [Dung, 1995] defines posterior (or conditional) probabilities of acceptability statuses of individual arguments given an attack relation between the same. The Bayesian network model then treats the acceptability statuses and attack relations as observable and unobservable data, respectively. As a result, the application of Bayes' theorem estimates the existence of attack relations with respect to unobserved relations by computing their posterior probabilities given observed acceptability statuses. The accuracy of correctness of the Bayesian network model is demonstrated by analysing properties of estimated attack relations and by illustrating the applicability of the proposed model to online forums.

The contribution of this study includes the following. To the best of the authors' knowledge, extant studies did not explore the use of acceptability semantics for statistical estimation of an attack relation. Dung's semantics is a classic theory of argument-based reasoning and Bayes' theorem is a classic

theory of statistical reasoning. Therefore, the present study corresponds to a milestone that allows various advanced studies that follow Dung to present their findings in the context of statistical reasoning. Nevertheless, the limitation of this study is that it is beyond the scope of the study to empirically discuss as to whether estimated attack relations are compelling to individuals in practice. For the sake of theoretical justification that motivates further empirical analysis, the study focuses on determination of properties of estimated attack relations under specific idealised and restricted conditions.

The study is organised as follows. Section 2 illustrates the idea of estimation of an attack relation. Section 3 describes background knowledge of computational argumentation and Bayesian inference. Section 4 shows the manner in which computational argumentation provides substance to Bayesian networks. Section 5 analyses the accuracy of the model, and finally Section 6 presents the conclusions and discusses related work.

2 Motivating Example

This section offers a simple example to demonstrate how the estimation of attack relations can be described as a Bayesian inference. We consider the analysis of attack relations between two arguments, a and b, resulting in four hypothetical attack relations: $\emptyset$, $\{(a,b)\}$, $\{(b,a)\}$ and $\{(a,b),(b,a)\}$ where (x,y) denotes that argument x attacks argument y. This yields the following four directed graphs (or abstract argumentation frameworks (AFs)): $AF_1 = \langle\{a,b\},\emptyset\rangle$, $AF_2 = \langle\{a,b\},\{(a,b)\}\rangle$, $AF_3 = \langle\{a,b\},\{(b,a)\}\rangle$ and $AF_4 = \langle\{a,b\},\{(a,b),(b,a)\}\rangle$, where each node and edge represents an argument and an attack relation between arguments, respectively.

According to the acceptability semantics [Dung, 1995], the acceptability status of an argument is interpreted differently for each AF. For example, $\{a,b\}$ denotes the complete extension[1] of AF_1 because no argument is attacked. $\{a\}$ (resp. $\{b\}$) denotes the complete extension of AF_2 (resp. AF_3) because only b (resp. a) is attacked. Finally, $\emptyset$, $\{a\}$ and $\{b\}$ denote the complete extensions of AF_4 because the symmetric attack relation between a and b results in three possible interpretations: the first interpretation is that neither argument is acceptable; the second interpretation is that only argument a is acceptable; and the third interpretation is that only argument b is acceptable.

Figure 1 shows the acceptability statuses of the arguments in each AF. Each of the four outside boxes represents an AF, and each of the inside boxes represents a complete extension defined in the AF. Each of the two circles within an inside box represents an acceptability status of arguments defined in the extension. Additionally, x and $\neg x$ denote that "argument x is acceptable" and "argument x is not acceptable", respectively.

We next consider an agent for which both arguments are observed, the casting of vote against a, and the possibility of identifying an attack relation between the two arguments from this observation. A novelty of our study is to consider

[1]Intuitively, an extension is a set of acceptable arguments, defined as a set of arguments such that its elements are defended by that set. For formal definitions see Section 3.

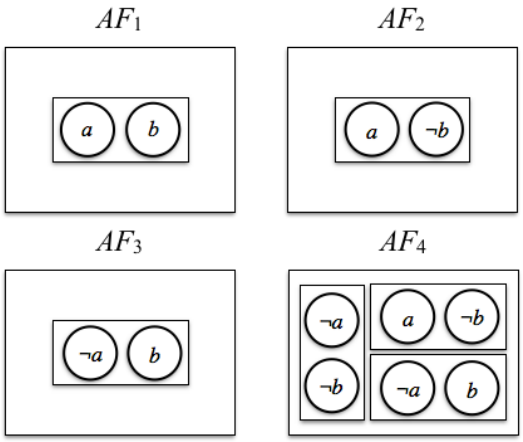

Figure 1: Abstract argumentation frameworks represented as outside boxes, their complete extensions as inside boxes and their acceptability statuses of arguments as circles.

that the voting agent has a particular AF and that she votes in accordance with the acceptability semantics. Therefore, the agent's vote can be considered as the process of selecting a ball $\neg a$ from a box in the AF. Intuitively, it is plausible that $\neg a$ comes neither from either AF_1 or AF_2 and rather from AF_3 or AF_4. In addition, we consider a further scenario of observing 100 votes against both arguments, a and b. Obviously, these votes make only AF_4 probable. In order to formalise these intuitions to discuss AFs these observations come from, this paper introduces Bayesian inference to calculate posterior probabilities for attack relations given votes.

3 Preliminaries

3.1 Computational Argumentation

An abstract argumentation framework (AF) [Dung, 1995] is defined as 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 iff (i.e., if and only if) some member of S attacks a. S is conflict-free iff S attacks none of its members. S defends a iff S is conflict-free and S attacks all arguments that attack a. A characteristic function $F : Pow(Arg) \to Pow(Arg)$ is defined by $F(S) = \{a | S \text{ defends } a\}$. Given AF, the acceptability semantics [Dung, 1995] defines four types of extensions that correspond to intuitively rational sets of arguments. S is a complete extension iff S is a fixed point of F. S is a grounded extension iff it is the minimum complete extension with respect to set inclusion. S is a preferred extension iff it is a maximal complete extension with respect to set inclusion. S is a stable extension iff it is a complete extension that attacks all members in $Arg \setminus S$.

Example 1. *It is assumed that $AF = \langle Arg, Att\rangle$ denotes an abstract argumentation framework where $Arg = \{a,b,c,d\}$ and $Att = \{(b,c),(c,b),(c,d),(d,d)\}$. Dung's acceptability semantics results in the following four types of extensions.*

- *Preferred extensions:* $\{a,b\}, \{a,c\}$

- *Stable extension:* $\{a, c\}$
- *Grounded extension:* $\{a\}$
- *Complete extensions:* $\{a\}, \{a, b\}, \{a, c\}$

A propositional language is introduced to define possible logical expressions of acceptability statuses of arguments.

Definition 1 (Language)**.** *A propositional language L_{Arg} associated with Arg is defined as follows. For all arguments $x \in Arg$, x is a formula of L_{Arg}. When x and y are formulas of L_{Arg}, $(x \wedge y)$, $(x \vee y)$, $(x \rightarrow y)$ and $\neg x$ are formulas of L_{Arg}.*

In this study, $s(AF), p(AF), g(AF)$ and $c(AF)$ denote the sets of all stable, preferred, grounded and complete extensions of AF, respectively.

3.2 Bayesian Inference

For any random variables V, $dom(V)$ represents the domain of V. The lowercase v represents a specific value in $dom(V)$. The bold uppercase $\boldsymbol{V}$ represents a sequence $[V_1, V_2, ..., V_n]$ of random variables V_i, and the bold lowercase $\boldsymbol{v}$ represents a sequence $[v_1, v_2, ..., v_n]$ of specific values in $dom(V)$. Additionally, $P(V = v)$ (or simply $P(v)$) represents the probability that the random variable V takes the value v. Furthermore, $\boldsymbol{P}(V)$ represents a sequence $[P(V = v_1), P(V = v_2), ..., P(V = v_n)]$ of the probabilities that the random variable V takes each value $v_i \in dom(V)$. Three types of random variables $H, D_i (1 \leq i \leq n)$, and X are assumed where H represents a directly unobservable hypothesis, each D_i represents an observed datum, and X represents an observable although not yet observed (i.e., unobserved) datum. Bayes' theorem is given as follows:

$$\boldsymbol{P}(H|\boldsymbol{D}) = \frac{\boldsymbol{P}(\boldsymbol{D}|H)\boldsymbol{P}(H)}{\boldsymbol{P}(\boldsymbol{D})}$$

It is used for hypothesis estimation and data prediction. With respect to the estimation, Bayesian inference computes posterior probabilities of all hypotheses given observed data.[2] It is assumed that observed data are i.i.d (independent and identically distributed). Subsequently, the posterior probability of hypothesis h_i given observed data $\boldsymbol{d}$ is calculated as follows:

$$\begin{aligned} P(h_i|\boldsymbol{d}) &= \frac{P(\boldsymbol{d}|h_i)P(h_i)}{P(\boldsymbol{d})} \\ &= \frac{\prod_j P(d_j|h_i)P(h_i)}{P(\boldsymbol{d})} \end{aligned}$$

With respect to the prediction, Bayesian inference uses all hypotheses to compute the posterior probability of an unobserved datum given observed data. It is assumed that each hypothesis determines the joint probability distribution of unobserved data. The posterior probability of unobserved datum

[2]MAP (Maximum a posteriori) estimate computes only hypotheses with the highest posterior probability. Thus, it uses a best hypothesis while predicting unobserved data.

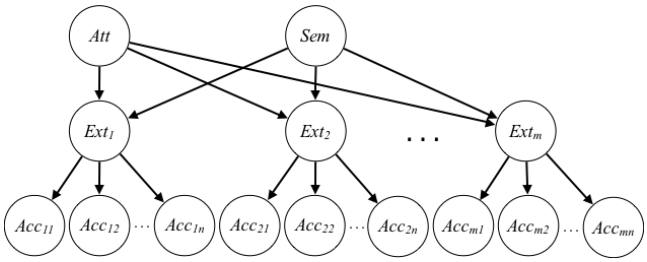

Figure 2: Bayesian network structure for argument-based reasoning.

x_i given observed data $\boldsymbol{d}$ is then calculated by the marginalisation of all possible hypotheses h_j.

$$\begin{aligned} x &= \arg\max_{x_i} P(x_i|\boldsymbol{d}) \\ &= \arg\max_{x_i} \sum_j P(x_i|h_j)P(h_j|\boldsymbol{d}) \end{aligned}$$

A Bayesian network is a directed acyclic graph where each node and edge represents a random variable and an independence relation between random variables, respectively. Each node V has a conditional probability distribution $\boldsymbol{P}(V|Parents(V))$ for the random variable V in which $Parents(V)$ denotes the set of random variables from which an edge exists to V. A full joint probability distribution for all random variables $\boldsymbol{V}$ in a Bayesian network is calculated as follows:

$$\boldsymbol{P}(\boldsymbol{V}) = \prod_{V \in \boldsymbol{V}} \boldsymbol{P}(V|Parents(V))$$

4 Estimation of Attack Relations

Four types of random variables Att, Sem, Ext, and Acc are assumed to represent attack relations, acceptability semantics, extensions, and acceptability statuses, respectively. It is assumed that Arg represents a set of arguments. The domain of Att is defined as a set of binary relations on Arg, i.e., $dom(Att) \subseteq Pow(Arg \times Arg)$, and the domain of Sem is defined as a set of the acceptability semantics, i.e., $Sem \subseteq \{s, p, g, c\}$ in which s, p, g, and c represents stable, preferred, grounded, and complete semantics, respectively. The domain of Ext is defined as a subset of the power set of arguments, i.e., $dom(Ext) \subseteq Pow(Arg)$, and the domain of Acc is defined as a set of a formula and its negation, i.e., $dom(Acc) = \{x, \neg x\}$ where $x \in L_{Arg}$. The dependencies among the random variables are defined as follows:

Definition 2 (Bayesian network structure)**.** *Let Att, Sem, Ext_i, and Acc_{ij} be random variables of attack relations, semantics, extensions, and acceptability statuses, respectively, for all $i (1 \leq i \leq m)$ and $j (1 \leq j \leq n)$. A Bayesian network for argument-based reasoning has the structure shown in Figure 2.*

Bold letters $\boldsymbol{Acc}$ are used to represent the sequence of sets $\{Acc_{11}, Acc_{12}, ..., Acc_{1n}\}$, $\{Acc_{21}, Acc_{22}, ..., Acc_{2n}\}$, ..., and $\{Acc_{m1}, Acc_{m2}, ..., Acc_{mn}\}$, i.e., $\boldsymbol{Acc} = [\{Acc_{ij} | 1 \leq j \leq n\} | 1 \leq i \leq m]$. This is followed by defining (un)conditional probabilities of the random variables. Given two randomly

chosen arguments, it is reasonable to assume that the possibility in which an argument attacks the other argument is lower than the possibility where this is not the case. This is because they are generally irrelevant. Thus, a higher probability is assigned to an attack relation when it involves a relatively small number of elements.[3]

Definition 3 (Prior probability of attack relations)**.** *Let att_i be an attack relation in which $1 \leq i \leq n$. The prior probability of att_i is defined as follows:*

$$P(Att = att_i) = \frac{1/(|att_i| + 1)}{\sum_{i=1}^{n}(1/(|att_i| + 1))}.$$

It should be noted that $\sum_{i=1}^{n} 1/(|att_i| + 1)$ denotes a constant. Specifically, it is viewed as a normalisation constant for the distribution $\boldsymbol{P}(Att)$. With respect to a prior probability of acceptability semantics, different semantics denote the different attitudes of an agent for the acceptance of arguments. In this study, it is assumed that each semantics occurs with the same probability.

Definition 4 (Prior probability of acceptability semantics)**.** *Let sem be an acceptability semantics. The prior probability of sem is defined by $P(Sem = sem) = 1/|dom(Sem)|$.*

A set of extensions is uniquely decided for an attack relation, and a semantics are determined. An agent's choice of an extension shows its preference for an outcome of argumentation. In this study, it is assumed that each extension occurs with the same probability.

Definition 5 (Posterior probability of extensions)**.** *Let ext be an extension, att be an attack relation, and sem be an acceptability semantics. The posterior probability of ext given att and sem is defined as follows:*

$$\begin{aligned} & P(Ext = ext | Att = att, Sem = sem) \\ = & \begin{cases} 1/|sem(\langle Arg, att\rangle)| & ext \in sem(\langle Arg, att\rangle) \\ 0 & (otherwise). \end{cases} \end{aligned}$$

Given an extension, an acceptability status of each argument is uniquely determined. Intuitively, it is necessary to define the posterior probability of its logical expression such that it corresponds to 1 if and only if the extension satisfies the formula in terms of the entailment relation $\models$. However, this can cause a zero-frequency problem in which a posterior probability of a dependent variable corresponds to 0 when only one formula that is not satisfied by the extension is observed. Thus, an m-estimator is used where m samples are assumed, and a few of the samples are satisfied and other samples are not satisfied by the extension. It is assumed that each of the m samples occurs with the same proportion p.

Definition 6 (Posterior probability of acceptability statuses)**.** *Let ext be an extension and acc be an acceptability status.*

[3]Various definitions are possible. For example, the same probability is assigned to each attack relation, a lower prior probability is assigned to asymmetric attack relations, or/and a lower prior probability is assigned to attack relations that form an odd loop.

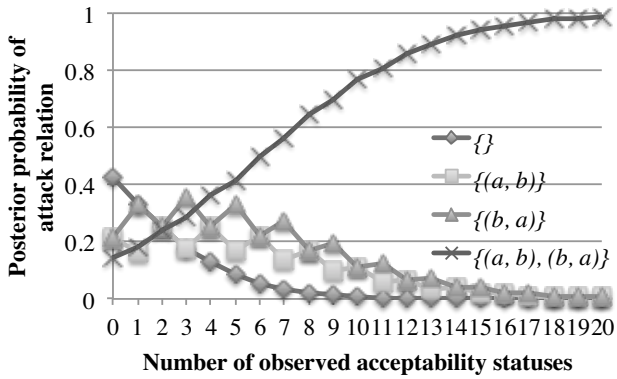

Figure 3: Posterior probabilities of the four attack relations given observed acceptability statuses.

The posterior probability of acc given ext is defined as follows:

$$\begin{aligned} & P(Acc = acc | Ext = ext) \\ = & \begin{cases} (1 + mp)/(1 + m) & ext \models acc \\ mp/(1 + m) & (otherwise). \end{cases} \end{aligned}$$

When m equals 2 and p equals 0.5, it is termed as a Laplace estimator that assumes two formulae acc and $\neg acc$ such that one formula is satisfied by the extension and the other formula is not. It considers the probability that acc is selected from the original one plus acc and $\neg acc$. As a result, $P(Acc = acc|Ext = ext) = 2/3$ if $ext \models acc$ otherwise $1/3$.

Example 2. *The following domains of random variables are considered.*

$$\begin{aligned} dom(Att) &= \{\emptyset, \{(a,b)\}, \{(b,a)\}, \{(a,b),(b,a)\}\} \\ dom(Sem) &= \{s, p, g, c\} \\ dom(Ext_i) &= \{\emptyset, \{a\}, \{b\}, \{a,b\}\} \\ dom(Acc_{i1}) &= \begin{cases} \{a, \neg a\} & \text{if } i \text{ is odd,} \\ \{b, \neg b\} & \text{if } i \text{ is even.} \end{cases} \end{aligned}$$

A situation in which the sequence $\boldsymbol{acc} = [\{\neg a\}, \{\neg b\}, \{\neg a\}, \{\neg b\}, \cdots, \{\neg a\}, \{\neg b\}]$ of acceptability statuses is observed is considered where $\neg a$ and $\neg b$ appear 10 times alternately and respectively. Figure 3 shows the posterior probabilities of attack relations with changes in the number of the observations. Their prior probabilities (i.e., when the number of data corresponds to 0) are defined based on the Definition 3, i.e., $P(\emptyset) = 3/7$, $P(\{(a,b)\}) = P(\{(b,a)\}) = 3/14$, $P(\{(a,b),(b,a)\}) = 1/7$. It is observed that the posterior probability of $\{(a,b),(b,a)\}$ converges to 1 with an increase in the observations.

5 Correctness

5.1 General Properties of Estimation

At least two strategies exist to evaluate the accuracy of the proposal. The first strategy corresponds to an empirical

evaluation to evaluate as to whether estimated attack relations are compelling to individuals in practice. The second strategy corresponds to a theoretical evaluation to evaluate as to whether the proposed Bayesian network provides expected attack relations under various conditions. This study focuses on the theoretical evaluation because it makes it clear that the proposed Bayesian network merits the empirical evaluation. In the following section, it is assumed that $dom(Att) = Pow(Arg \times Arg)$, $dom(Sem) = \{s, p, g, c\}$, $dom(Ext) = Pow(Arg)$, and $dom(Acc) = \{x, \neg x\}$ where $x \in L_{Arg}$. Moreover, it is assumed that Proposition 1 and Corollary 1 assume that an m-estimator does not exist, and thus it is assumed that $m = p = 0$ in Definition 6.

Proposition 1. *Let $\boldsymbol{acc}$ be a sequence of acceptability statuses, and att be an attack relation. $P(att|\boldsymbol{acc}) > 0$ iff, for all $acc \in \boldsymbol{acc}$, there exist a semantics $\varepsilon \in \{g, p, s, c\}$ and an extension $E \in \varepsilon(\langle Arg, att\rangle)$ such that $E \models acc$.*

The following corollary states that an estimation is impossible when an anomaly is observed.

Corollary 1. *Let $\boldsymbol{acc}$ be a sequence of acceptability statuses. If there is $acc \in \boldsymbol{acc}$ such that acc is not satisfiable, then $P(att|\boldsymbol{acc}) = 0$ holds for all attack relations att but not vice versa.*

The corollary holds because no extension satisfies any unsatisfiable set of formulas. It should be noted that such an anomaly is successfully handled under the m-estimator. The next proposition states that an estimation is useless when an obvious observation is noted.

Proposition 2. *Let $\boldsymbol{acc}$ be a sequence of acceptability statuses. If $\boldsymbol{acc}$ is a sequence of sets of tautologies, then $P(att|\boldsymbol{acc}) = P(att)$ holds for all attack relations att but not vice versa.*

This proposition holds because every extension satisfies any tautology that provides no information for the estimation. The next proposition concerns a limitation of the estimation.

Proposition 3. *Let $\boldsymbol{acc}$ be a sequence of acceptability statuses. For all attack relations att_1 and att_2, and acceptability semantics $\varepsilon \in \{c, p, s, g\}$, $P(att_1|\boldsymbol{acc}) \geq P(att_2|\boldsymbol{acc})$ if $\varepsilon(\langle Arg, att_1\rangle) = \varepsilon(\langle Arg, att_2\rangle)$ and $P(att_1) \geq P(att_2)$.*

A positive interpretation of the proposition is that the estimation conforms to the extensions. A negative interpretation is that the estimation cannot distinguish different attack relations that result in the same extensions. The next corollary directly follows from the proposition.

Corollary 2. *Let $\boldsymbol{acc}$ be a sequence of acceptability statuses. For all attack relations att_1 and att_2, and acceptability semantics $\varepsilon \in \{c, p, s, g\}$, $P(att_1|\boldsymbol{acc}) = P(att_2|\boldsymbol{acc})$ if $\varepsilon(\langle Arg, att_1\rangle) = \varepsilon(\langle Arg, att_2\rangle)$ and $P(att_1) = P(att_2)$.*

5.2 Estimation Under Data Restriction

In the previous subsection, no restriction is imposed on observed acceptability statuses. In this subsection, they are idealised and an investigation is conducted as to whether the estimation provides expected attack relations. The following theorem states that when acceptability statuses correspond to extensions of an abstract argumentation framework, the estimated attack relations result in the same extensions that are provided by the framework.

Theorem 1. *For all binary attack relations att^* on Arg and acceptability semantics $\varepsilon \in \{c, p, s, g\}$, the following relation holds:*

$$\varepsilon(\langle Arg, att^*\rangle) = \varepsilon(\langle Arg, \arg\max_{att} P(att|\boldsymbol{acc})\rangle),$$

where $\boldsymbol{acc}$ denotes the sequence of the sets $E \cup \{\neg x | x \notin E\}$, for all extensions $E \in \varepsilon(\langle Arg, att^*\rangle)$.

Proof. Let $\hat{att}$ denote an estimated attack relation. It is shown that if $\varepsilon(\langle Arg, att^*\rangle) \neq \varepsilon(\langle Arg, \hat{att}\rangle)$ holds then $\hat{att} \neq \arg\max_{att} P(att|\boldsymbol{acc})$ holds. Let $\boldsymbol{acc}$ denote [{acc_{11},..., acc_{1n}},..., {acc_{m1},..., acc_{mn}}]. $P(att|\boldsymbol{acc})$ is given as follows:

$$\alpha P(att|\boldsymbol{acc}) = \alpha P(att) \sum_{sem} P(sem) \prod_{i=1}^{m} \sum_{ext_i} P(ext_i|att, sem) \prod_{j=1}^{n} P(acc_{ij}|ext_i),$$

where α denotes the normalisation constant. Given the assumption on $\boldsymbol{acc}$, for all acc_{ij}, there exists a $ext_i \in \varepsilon(\langle Arg, att^*\rangle)$ such that $ext_i \models acc_{ij}$ holds. However, this is not the case with respect to $ext_i \in \varepsilon(\langle Arg, \hat{att}\rangle)$. Therefore, $P(att^*|\boldsymbol{acc}) > P(\hat{att}|\boldsymbol{acc})$ (and $P(\hat{att}|\boldsymbol{acc}) = 0$ when $m = p = 0$ holds in m-estimator) holds. □

A positive interpretation of the theorem is that the Bayesian network provides reasonable attack relations in the sense that the estimated attack relations result in the same extensions as the true attack relation. A negative interpretation is that the Bayesian network is generally unable to detect the true attack relation even when acceptability statuses are idealised. However, the following theorem shows that a certain restriction on attack relations overcomes this limitation.

Theorem 2. *For all symmetric and irreflexive attack relations att^* on Arg and acceptability semantics $\varepsilon \in \{c, p, s\}$ except grounded semantics g, the following relation holds:*

$$att^* = \arg\max_{att} P(att|\boldsymbol{acc}),$$

where $\boldsymbol{acc}$ denotes the sequence of the sets $E \cup \{\neg x | x \notin E\}$, for all extensions $E \in \varepsilon(\langle Arg, att^*\rangle)$.

Proof. This directly follows from the fact that for any two symmetric and irreflexive abstract argumentation frameworks AF_1 and AF_2 with the same set of arguments, if $\varepsilon(AF_1) = \varepsilon(AF_1)$ then $AF_1 = AF_2$ for all $\varepsilon\{c, p, s\}$. Therefore, Theorem 1 guarantees that the estimation results in att^*. □

The theorem states that when acceptability statuses are extensions of an abstract argumentation framework with a symmetric and irreflexive attack relation, then the estimated attack relation is the same as the framework. However, this theorem does not hold under the grounded semantics. For example, the proof of Theorem 2 does not

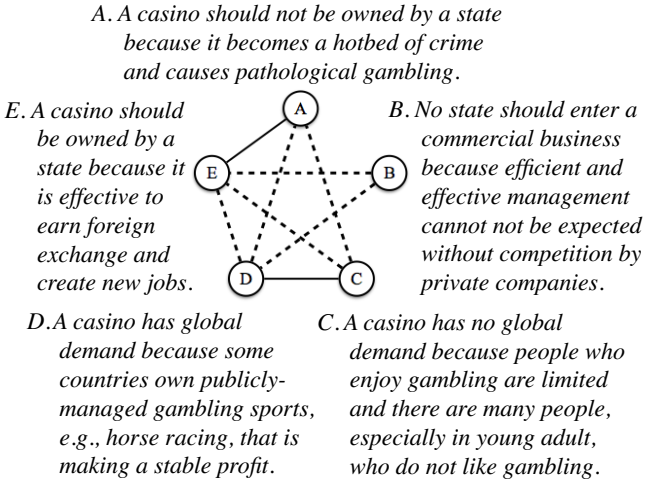

Figure 4: Argument on a government-controlled casino.

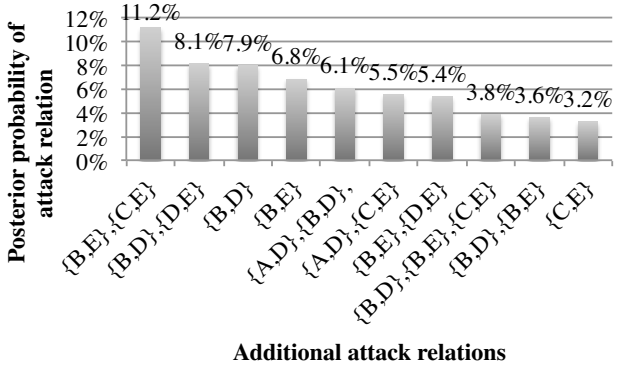

Figure 5: Top 10 posterior probabilities of additional attack relations given the acceptability statuses from every agent.

hold when the grounded semantics is applied to the following: $AF_1 = \langle\{a, b, c\}, \{(a, b), (b, a), (b, c), (c, b)\}\rangle$ and $AF_2 = \langle\{a, b, c\}, \{(a, b), (b, a), (a, c), (c, a)\}\rangle$. Specifically, the grounded semantics is subject to Theorem 1.

5.3 Illustration of Its Applicability

The aim of this subsection involves demonstrating the accuracy of this research in terms of its applicability to online forums. The study illustrates and discusses using arguments and their votes to estimate attack relations. A situation is considered in which individuals argue for and against a government-controlled casino in an online forum. Figure 4 shows five arguments put forward by the individuals. Arguments connected by the solid line (resp. no line) represent the fact that individuals acknowledge that there is a symmetric (resp. no) attack relation between the same. Arguments connected by the dotted line represent that the existence of attack relations is not clear, and is therefore the subject of estimation.

Table 1 shows the votes of 20 agents for and against each of the five arguments. Each $+$ and $-$ represents a positive and negative vote, respectively, and the blanks represent missing values. When each vote is linked to an agent, it is rational to assume that votes by the same agent occur from the same extension of the same abstract argumentation framework. Moreover, it is rational to represent each $+$ and $-$ by an atomic formula and its negation of L_{Arg}, respectively, where $Arg = \{a, b, c, d, e\}$. Therefore, the following sequence $\boldsymbol{acc}$ of acceptability statuses is obtained from every agent:

$$\boldsymbol{acc} = [\{a, \neg b, c, \neg d, \neg e\}, \{\neg a, b, \neg c, d, e\}, ..., \{\neg a, \neg b, e\}],$$

where each set in the sequence corresponds to each agent's votes. Conversely, when each vote is not linked to an agent, another method involves focusing on the sums of votes for and against each argument as shown in the right ends of Table 1. It is now reasonable to assume that each vote might not occur from the same extension. Therefore, the following sequence of acceptability statuses is obtained from every vote:

$$\boldsymbol{acc} = [\{a\}, ..., \{a\}, \{\neg a\}, ..., \{\neg a\}, ..., \{\neg e\}, ..., \{\neg e\}],$$

where each $\{x\}$ and $\{\neg x\}$ appears equal to the number of positive $(+)$ and negative $(-)$ votes, respectively, for all arguments X.

With respect to the acceptability statuses from every agent, the top 10 estimated attack relations are shown in Figure 5 where for all arguments X and Y, $\{X, Y\}$ corresponds to the abbreviation of $\{(X, Y), (Y, X)\}$. For the sake of clarity, only symmetric attack relations without an odd loop are considered in the estimation.

The acceptability statuses discussed in the two examples are both expressed by literals of the propositional language. However, it should be noted that the proposed system can handle statuses expressed by any propositional formulae.

6 Conclusions and Discussion

In this study, a Bayesian approach to argument-based reasoning is proposed for statistically estimating the existence of an attack relation existing between arguments. Dung's acceptability semantics is utilised to infer attack relations from acceptability statuses of arguments. The results indicate that the proposed Bayesian network provides reasonable attack relations under idealised and restricted conditions. The applicability of the proposed Bayesian network is illustrated by showing the manner in which votes by agents in online forums are used to estimate attack relations.

There are extant studies related to computational argumentation to Bayesian inference. Previous studies in these fields can be considered from at least three directions. The aim of the first direction involves examining decision making frameworks for multi-agent systems. For example, Nielsen and Parsons [Nielsen and Parsons, 2007] deal with a fusion of Bayesian networks in multi-agent systems. They provide a framework governed by principles of formal argumentation that allows agents to finally agree on a single Bayesian network. Saha and Sen [Saha and Sen, 2004] use a Bayesian network to provide a model of the mental states of agents in

Table 1: Votes by 20 agents for (denoted by $+$) or against (denoted by $-$) 5 arguments.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Sum: $+$	$-$
A	$+$	$-$	$+$	$+$	$+$	$-$	$+$	$+$		$+$	$-$	$+$		$+$	$-$	$-$		$+$		$-$	10	6
B	$-$	$+$	$+$	$-$	$-$	$-$	$+$	$-$	$-$	$-$		$+$	$+$	$+$		$+$	$+$	$+$	$-$	$-$	9	9
C	$+$	$-$	$+$	$+$	$-$	$-$		$+$	$+$	$-$				$+$					$-$		6	5
D	$-$	$+$	$-$	$+$	$+$	$-$		$+$	$+$	$-$	$-$	$-$		$+$		$-$		$-$			6	8
E	$-$	$+$	$-$	$-$	$+$	$-$	$-$	$+$		$+$	$-$	$-$	$+$	$+$	$-$	$+$		$-$		$+$	8	9

argument-based automated negotiation. The aim of the second direction involves investigating the conversion of probabilistic domain knowledge from Bayesian networks to structured arguments. For example, Timmer et al. [Timmer *et al.*, 2015] attempted to incorporate probability into models of argumentation and proposed a method to build arguments from Bayesian networks. Vreeswijk [Vreeswijk, 2005] discussed a method to extract arguments and attacks from domain knowledge represented by a Bayesian network. The aim of the third direction involves examining the conversion of probabilistic domain knowledge from structured arguments to Bayesian networks. For example, Grabmair et al. [Grabmair *et al.*, 2010] provided a translation of Carneades models of argumentation into a Bayesian network. Bex and Renooji [Bex and Renooij, 2016] proposed a method of deriving constraints on Bayesian network based on argument-based reasoning.

In contrast, the aim of the present study involves identifying an attack relation, i.e., one component of argumentation. The aim does not involve providing a framework for organising various components of argumentation. Moreover, the study provides a Bayesian network model of argument-based reasoning albeit without the aim of providing a model of representing domain knowledge used in argument-based reasoning.

Acknowledgements

The authors thank Prof. Shier Ju for his kind support and Prof. Ken Satoh for valuable discussion. This study was supported by JSPS KAKENHI Grant Number 15KT0041.

References

[Bex and Renooij, 2016] Floris Bex and Silja Renooij. 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, 2016.

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

[Dunne *et al.*, 2015] Paul E. Dunne, Wolfgang Dvořák, Thomas Linsbichler, and Stefan Woltran. Characteristics of multiple viewpoints in abstract argumentation. *Artificial Intelligence*, 228:153–178, 2015.

[Grabmair *et al.*, 2010] Matthias Grabmair, Thomas F. Gordon, and Douglas Walton. 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, 2010.

[Lawrence and Reed, 2016] John Lawrence and Chris Reed. Argument mining using argumentation scheme structures. In *Proc. of the 6th International Conference on Computational Models of Argument*, pages 379–390, 2016.

[Moens, 2013] Marie-Francine Moens. Argumentation mining: Where are we now, where do we want to be and how do we get there? In *Proc. of the 5th Forum on Information Retrieval Evaluation*, 2013.

[Nielsen and Parsons, 2007] Søren Holbech Nielsen and Simon Parsons. An application of formal argumentation: Fusing bayesian networks in multi-agent systems. *Artificial Intelligence*, 171:754–775, 2007.

[Niskanen *et al.*, 2016] Andreas Niskanen, Johannes P. Wallner, and Matti Järvisalo. Synthesizing argumentation frameworks from examples. In *Prof. of the 22nd European Conference on Artificial Intelligence*, pages 551–559, 2016.

[Palau and Moens, 2009] Raquel Mochales Palau and Marie-Francine Moens. 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, 2009.

[Riveret and Governatori, 2016] Régis Riveret and Guido Governatori. On learning attacks in probabilistic abstract argumentation. In *Proc. of the 15th International Conference on Autonomous Agents and Multiagent Systems*, pages 653–661, 2016.

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

[Timmer *et al.*, 2015] Sjoerd T. Timmer, John-Jules Ch. Meyer, Henry Prakken, Silja Renooij, and Bart Verheij. Explaining bayesian networks using argumentation. In *Proc. of the 13th European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 83–92, 2015.

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