Enumerating preferred extensions: A case study of human reasoning

Alice Toniolo[1], Timothy J. Norman[2], and Nir Oren[3]

[1] School of Computer Science, University of St Andrews, Scotland, UK
[2] Department of Electronics and Computer Science, University of Southampton, UK
[3] Department of Computing Science, University of Aberdeen, Scotland, UK

Abstract. This paper seeks to better understand the links between human reasoning and preferred extensions as found within formal argumentation, especially in the context of uncertainty. The degree of believability of a conclusion may be associated with the number of preferred extensions in which the conclusion is credulously accepted. We are interested in whether people agree with this evaluation. A set of experiments with human participants is presented to investigate the validity of such an association. Our results show that people tend to agree with the outcome of the probabilistic semantics in purely qualitative domains as well as in domains in which conclusions express event likelihood. Furthermore, we are able to characterise this behaviour: the heuristics employed by people in understanding preferred extensions are similar to those employed in understanding probabilities.

Keywords: Argumentation, Probabilistic Semantics, User Evaluation

1 Introduction

One of the strengths of argumentation theory is its qualitative nature. For example, in Dung's theory, arguments are either within, or outside an extension, and no notion of argument strength is required in order to obtain desirable features — such as reinstatement — from the system. More recently, researchers have begun considering more quantitative frameworks, particularly in the context of probabilistic argumentation (e.g., [8, 10, 11, 18]), through weighted argumentation systems [2, 7] and graduality within argumentation [4]. The immediate question then arises as to whether such quantitative representations appropriately capture human reasoning and intuitions, as well as questions regarding the relationship between formal qualitative representations and human quantitative (or semi-quantitative) reasoning. As a concrete example — which we focus on in this paper — one could view multiple extension semantics, such as the preferred semantics, as capturing different possible worlds. This would then suggest that even qualitative argumentation can capture some notion of uncertainty.

This view can be further extended by considering situations where the arguments within an extension are themselves about uncertain facts, effectively changing the likelihood of each extension. If this is the case, then even in purely qualitative domains (represented through logical argumentation), where no quantified information exists, the degree of acceptability of a conclusion is associated with the number of preferred extensions in which the conclusion is credulously accepted. This paper investigates the validity of this claim, by means of an experiment with human participants.

The remainder of the paper is structured as follows. In Section 2, we expand the motivations of this work. In Section 3, we introduce an ASPIC-like argumentation framework followed by an overview of its use and key assumptions underpinning our experiments (Section 4). Section 5 details our experimental settings. In Section 6, methodology, hypotheses and results are discussed. We present our conclusions in Section 7.

2 Background and motivation

Haenni [8] considers uncertainty as being an evaluation of probability on the premises which propagates throughout the argumentation system. Similarly, other studies such as [18] and [15] model uncertainty on the premises as being associated with the uncertainty of the sources, in the latter case due to the different degrees of trustworthiness of the sources themselves. Li et al [11] consider a different take on probability, namely that the probability of an argument represents a prediction on how likely it is that the argument is justified.

In this work, we are interested in studying the links between the preferred extensions as used in argumentation, and how these are interpreted as probabilities by people with regards to the acceptability of a conclusion. Let us consider a conclusion of an argument within a structured argumentation framework. Generally, argumentation frameworks presented in the literature use extensions to decide whether a conclusion is accepted. In purely qualitative argumentation frameworks, this acceptance is either credulous (when there is at least one extension in which the argument under consideration is accepted), or sceptical (when the argument is accepted in all extensions) [13]. As dictated by the nature of qualitative frameworks, the enumeration of extensions in which a conclusion is accepted does not influence the decision as to whether a conclusion is accepted. However, here we claim that the number of extensions in which a conclusion is accepted has an effect on deciding whether the conclusion is to be considered justified, even if the argumentation framework is fully qualitative[4].

The problem of understanding the role of enumeration of extensions has been studied by Thimm [16] in abstract argumentation. Thimm presents a novel argumentation framework in which a probabilistic semantics is used to associate an argument with a degree of belief. This belief is computed as function of the number of extensions in which the argument appears to be justified. In our work, we use a similar approach where we consider the enumeration of preferred

[4] Note that we use the terms argument and conclusion somewhat interchangeably as in the work we describe, a specific conclusion was the result of a unique argument. As future work, we will consider situations where multiple arguments may lead to the same conclusion, c.f., the so called universal semantics [6].

extensions in evaluating the believability of a conclusion. Thimm claims that this assessment provides a degree of confidence when selecting an option. Here we want to understand whether this is the case, i.e., whether people do use a similar heuristic to make a decision on what conclusions are the most believable. In Thimm's work, a probability is associated with each extension, and this influences the degree of belief placed in an argument. In our study we want to understand whether doing so is comparable to human reasoning with probability.

Unlike Thimm's work, we use structured argumentation frameworks, as we are interested in the believability of conclusions rather than arguments. Our core research question is then as follows: *do people agree with the evaluation given by probabilistic interpretation of argumentation semantics?* To address this question, we define an ASPIC-like structured argumentation framework from which we can formalise the problem.

3 An ASPIC-like framework with probabilistic semantics

In order to identify plausible conclusions, we use a simplified ASPIC-like argumentation framework with ordinary premises and defeasible rules without preferences or undercuts [13, 14]. We derive the degree of belief in a conclusion obtained by applying argumentation semantics to arguments obtained from the framework, and considering a probabilistic interpretation of the results.

3.1 Argumentation framework

Definition 1. *An* argumentation system *AS is a tuple* $\langle \mathcal{L}, \bar{}, \mathcal{R} \rangle$ *where* $\mathcal{L}$ *is a logical language,* $\bar{}$ *is a contrariness function, and* $\mathcal{R}$ *is a set of defeasible rules. The contrariness function* $\bar{}$ *is defined from* $\mathcal{L}$ *to* $2^{\mathcal{L}}$*, such that given* $\varphi \in \bar{\phi}$ *with* $\varphi, \phi \in \mathcal{L}$*, if* $\phi \notin \bar{\varphi}$*,* φ *is called the* contrary *of* ϕ*, otherwise if* $\phi \in \bar{\varphi}$ *they are* contradictory *(including classical negation* $\neg$*). A* defeasible rule *is* $\varphi_0, \ldots, \varphi_j \Rightarrow \varphi_n$ *where* $\varphi_i \in \mathcal{L}$*.*

Definition 2. *A* knowledge-base *K in AS is a subset of the language* $\mathcal{L}$*. An* argumentation theory *is a pair* $AT = \langle K, AS \rangle$*.*

An *argument* A is derived from K of theory AT. Let $Prem(A)$ indicate the premises of A, $Conc(A)$ the conclusion, and $Sub(A)$ the subarguments:

Definition 3. *Given a set of arguments Arg, argument* $A \in Arg$ *is defined as:*

- $A = \{\varphi\}$ *with* $\varphi \in K$ *where* $Prem(A) = \{\varphi\}$*,* $Conc(A) = \varphi$*,* $Sub(A) = \{\varphi\}$*.*
- $A = \{A_1, \ldots, A_n \Rightarrow \phi\}$ *if there exists a defeasible rule in AS s.t.* $Conc(A_1), \ldots, Conc(A_n) \Rightarrow \phi \in \mathcal{R}$ *with* $Prem(A) = Prem(A_1) \cup \cdots \cup Prem(A_n)$*,* $Conc(A) = \phi$ *and* $Sub(A) = Sub(A_1) \cup \cdots \cup Sub(A_n) \cup A$*.*

Attacks are defined as those arguments that challenge others, while defeats are those attacks that succeed:

Definition 4. *Given two arguments* A_A *and* A_B*:*

- A_A rebuts A_B *on* $Arg_{B'}$ *iff* $Conc(A_A) \in \bar{\varphi}$ *for* $A_{B'} \in Sub(A_B)$ *such that* $A_{B'} = \{A_{B1"}, \ldots, A_{Bn"} \Rightarrow \varphi\}$.
- A_A undermines A_B *on* φ *iff* $Conc(A_A) \in \bar{\varphi}$ *such that* $\varphi \in Prem(A_B)$.

Definition 5. *Defeat is a binary relationship* $Def : Arg \times Arg$ *where a defeat is represented as* $(A_A, A_B) \in Def$*. An argument* A_A defeats *an argument* A_B *iff:* i) A_A *rebuts* A_B *on* $A_{B'}$*; or* ii) A_A *undermines* A_B *on* φ*.*

Definition 6. *An abstract argumentation framework* $AF = (Arg, Def)$ *corresponding to an AT contains the set of arguments Arg as defined in Definition 3 and a set of defeats Def as in Definition 5.*

Sets of acceptable arguments (i.e., extensions ξ) in an AF can be computed according to a semantics. Here we use the preferred semantics. The set of credulous preferred extensions is $\hat{\xi}_P = \{\xi_1, ..., \xi_n\}$, where every ξ_i is a maximal set of arguments (with respect to set inclusion) that is conflict free and admissible.

Definition 7. *Given an abstract argumentation framework* $AF = (Arg, Def)$*, a set of arguments* $S \subseteq Arg$ *is* conflict-free *iff there is no* $A_A, A_B \in S$ *such that* $(A_A, A_B) \in Def$*. An argument* $A_A \in S$ *is admissible iff for every* A_B *such that* $(A_B, A_A) \in Def$*, there is a* $A_C \in S$ *such that* $(A_C, A_B) \in Def$*.*

3.2 Probabilistic semantics for an argument theory

Having described a simple ASPIC-like framework, we now describe how Thimm's probabilistic semantics [16] is used to associate probabilities with conclusions.

The set of all possible sets of arguments is referred to as $\mathcal{K} = 2^{Arg}$, and the set of preferred extensions $\hat{\xi}_P$ is a subset of $\mathcal{K}$. A probability function of the form $P : 2^{\mathcal{K}} \to [0, 1]$ assigns to each set of possible extensions of AF a probability. For $\xi \in \mathcal{K}$, $P(\xi)$ is the probability that ξ is an extension. For now, we make the assumption that extensions are equiprobable. Then the probability of ξ is:

$$P(\xi) = \begin{cases} 1/|\hat{\xi}_P| & \xi \in \hat{\xi}_P \\ 0 & \xi \notin \hat{\xi}_P \end{cases} \tag{1}$$

For $P(\xi)$ and argument $A \in Arg$:

$$\hat{P}(A) = \sum_{A \in \xi \subseteq Arg} P(\xi) \tag{2}$$

Given the probability function P, $\hat{P}(A)$ represents the degree of belief that an argument A is in an extension according to P.

As Thimm suggests we now have an indication of the degree of belief of each argument that gives a characterisation of the uncertainty which is inherent in the AF. We must define several additional concepts in order to describe the acceptability of conclusions within the argumentation framework.

From [13] we know that a wff $\varphi \in \mathcal{L}$ is sceptically justified if φ is the conclusion of a sceptically justified argument, and credulously justified if φ is not sceptically justified and is the conclusion of a credulously justified argument. Hence we define a *justification ratio* μ of a conclusion φ as follows.

$K = \{r_1; r_2; r_3; r_4\}$ $\mathcal{R} = \{r_4 \Rightarrow r_5\}$

$^{-} = \{(r_2, r_3); (r_3, r_2); (r_2, r_1); (r_3, p_1); (r_5, r_1)\}$

Arguments:

$A_1 : r_4$ $A_2 : r_2$ $A_3 : r_3$ $A_4 : r_1$ $A_5 : A_1 \Rightarrow r_5$

Preferred extensions: $\xi_1 = \{A_1, A_3, A_5\}$ $\xi_2 = \{A_1, A_2, A_5\}$

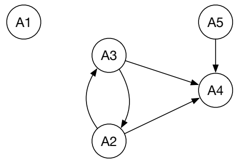

Fig. 1. Example of argumentation theory

Definition 8. *Given a set of arguments* $\mathcal{A} = \{A_1, \ldots, A_n\}$ *such that for any* A_i, $Conc(A_i) = \varphi$, *we define the* justification ratio *as* $\mu(\varphi) = \sum_{A_i \in \mathcal{A}} \hat{P}(A_i)$.

The justification ratio $\mu(\phi)$ captures the probability of a conclusion being justified based on the likelihood of the arguments which justify it. If equiprobable extensions are assumed, then we obtain:

$$\mu(\varphi) = \hat{P}(A) = \sum_{A \in \xi \subseteq Arg} 1/|\hat{\xi}_P| \quad \text{where } \varphi \in Conc(A)$$

Example 1. We now illustrate the framework with the following example. Consider the *AT* presented in Figure 1. We obtain two preferred extensions ξ_1, ξ_2 with $P(\xi_1) = P(\xi_2) = 0.5$. The justification ratios are then as follows:

$$\mu(r_1) = 0 \qquad \mu(r_2) = \mu(r_3) = 0.5 \qquad \mu(r_4) = \mu(r_5) = 1$$

4 Characterising reasoning with extensions

In the previous section, we explored a method to assign a degree of belief to a conclusion (which we denoted as the justification ratio) in relation to the enumeration of extensions by adapting Thimm's probabilistic semantics. Our main objective is to determine whether people agree with these probabilistic semantics; i.e., whether the justification ratio has a correlation with people's opinion of the believability of a conclusion. We believe that this is the case on the basis of the assumption that *people's reasoning with extensions may be understood in relation to reasoning with the rules of classical probability.* This assumption serves us as second objective, i.e., that of characterising how people rate the believability of a conclusion.

We now discuss the observations underpinning our assumption. Consider as a domain the scenario of social inferences — inferences drawn from social media information and corroborated with background knowledge to draw potentially unwanted conclusions [12]. Assume that an argumentation framework is used to decide whether some conclusions can be inferred from a set of information previously shared within a social network. If these conclusions are justified within one or all of the extensions, these can be considered more likely to be inferred (and hence known) by users of the social network. Similarly consider a set of information and associated likelihood of this information being shared in the social network. The more likely it is that a piece of information is shared, the more likely is that this will be known by some users in the network.

In more formal terms, our analysis is based on the following observations:

- Classical probability assigns the likelihood to a piece of information φ on the basis of the ratio of the number of cases favourable to an event to unfavourable cases. In the above example, the event may be considered as that of the information being shared. Hence, consider a set of possible worlds W and a subset of the worlds $V \subset W$ in which a proposition $r_i \in \mathcal{L}$ holds, the probability of r_i is as follows.

$$p(r_i) = \frac{\#\text{ of worlds where } r_i \text{ holds}}{\text{total } \# \text{ possible worlds}} = \frac{|V|}{|W|}$$

- Similarly, if we consider the set of preferred extensions $\hat{\xi}_P$ as the set of possible explanations of a world, and the degree of belief of a conclusion r_i is given by the justification ratio $\mu(r_i)$. Let us refer to the subset of extensions in which r_i is acceptable as $\hat{\xi}_P^{r_i}$. From Definition 8 we obtain the following.

$$\mu(r_i) = \sum_{A \in \xi \subseteq Arg} 1/|\hat{\xi}_P| = \frac{\#\text{ extensions in which } r_i \text{ is acceptable}}{\text{total } \# \text{ extensions}} = \frac{|\hat{\xi}_P^{r_i}|}{|\hat{\xi}_P|}$$

In the above situation, we assume that the information is purely qualitative. However, the information may refer to the likelihood of an event or a fact [1]. For example, an event E described in r_i can be subject of a proposition r_j= "there is a ω chance that event E may occur". Continuing with the similarity between reasoning with extensions and reasoning with probability, we also seek to understand the behaviour in the case in which the user is presented with information that is about the likelihood of events, as well as the uncertainty introduced via the possibility of some information being or not being inferred. In this case, the believability of a conclusion may be explained by two heuristics depending on whether people consider these as dependent or independent events. The similarity with an argumentation framework outcome can then be established in the former case through the use of conditional probability, or in the latter by using the multiplication law of probability. For this research, we assume that the second heuristic is adopted, resulting in the following observations:

- ω indicates the probability of the event $p(E)$. Given $p(r_i)$, the probability of r_j using the multiplication law for independent events is: $p(r_j) = \omega * p(r_i)$
- Similarly, in an argumentation framework with probabilistic semantics, given the justification ratio $\mu(r_i)$, the justification ratio of r_j is: $\mu(r_j) = \omega * \mu(r_i)$

We are now in a position to describe our experiments, designed to determine (1) whether the probabilistic interpretation described above represents human reasoning, and (2) whether the similarities observed between probabilistic and argument based reasoning are valid.

5 Experiment Design

Our overall objective is to understand whether people agree with the outcome of probabilistic semantics. In our experiments, we asked a participant to rate the believability of a propositions under different experimental conditions α. While considering different experimental conditions, we posed the following question to our subjects: "Given the condition α, how likely is that you believe r_i"? The subjects were asked to respond on a 5-points Likert scale, a commonly used scale for user studies, recorded as user evaluation $u(r_i)$ of a conclusion r_i (with 1: Extremely Unlikely – 5: Extremely Likely). Our hypothesis is that there is a positive correlation between the user rating $u_\mu(r_i)$ and the justification ratio $\mu(r_i)$. We also hypothesise that there is a positive correlation between the user evaluation of the likelihood of a piece of information r_i — $u_p(r_i)$ — and its associated probability $p(r_i)$. Finally, we intend to show that there is a similarity between the two ratings $u_p(r_i)$ and $u_\mu(r_i)$.

Definition 9. *An experimental condition α can is a tuple $\alpha = \langle$ Domain, Scenario, Proposition, Interpretation, Percentage, Fraction, Ratio$\rangle$.*

We now define the components of an experimental condition α.

5.1 Two Types of Information

As discussed in Section 4, information — represented via propositions — can be classified into two categories, or domain types in the context of the experiment.

Domain 1: Purely qualitative propositions $r_i \in \mathcal{L}$ in which the text is about a piece of information.

Domain 2: Propositions $r_j \in \mathcal{L}$ in which the text is about a piece of information and its probability of occurring.

In the former, we want to demonstrate that even in purely qualitative scenarios, people agree with the outcome of the probabilistic semantics: the believability of a conclusion is related to the number of extensions in which that conclusion is accepted. With the latter, we want to demonstrate that in scenarios in which conclusions are about the probability of some information, the outcome of the probabilistic semantics is still an important factor in assessing the believability of a conclusion. The two types of propositions lead to two sets of experiments.

5.2 Scenarios and Propositions

In the experiments we use seven scenarios within the social inference domain. While our work is generalisable to other domains, this seemed to lend itself well to the design of the experiments. The scenarios are derived from reported incidents in the context of sharing political views [9], and location data or temporal information [12]. These scenarios are built using a combination of arguments from position to know and cause to effect [17].

Each scenario is referred to as Xi with $i = 1$–7 and designed as a set of propositions $r_i \in \mathcal{L}$. In order to collect a relatively large amount of data with less cognitive effort for the user, two propositions per scenario are chosen as tested propositions subjects of our experiments. We combine proposition and scenario using the same notation, writing Xi_j, where $j = 0, 1$ refers to the proposition being tested. For convenience, we call Xi_j a scenario. Given 7 base scenarios and 2 propositions, we obtain a total of 14 scenarios.

5.3 Interpretations

For each scenario, two interpretations can be made:

At: An interpretation building on the number of extensions in which the conclusion is acceptable, and which considers the total number of extensions (via an argument theory AT, with rules and contraries between propositions), in which a justification ratio $\mu(r_i)$ is associated with each proposition r_i.

Pt: A possible worlds based probabilistic interpretation, in which each proposition r_i is associated with a probability $p(r_i)$ of its information being verified.

We associate the justification ratio of a conclusion r_i as the outcome of the probability semantics, with the likelihood that that piece of information is verified (e.g., is shared). Given that both interpretations are based on the same set of propositions, the key design link is such that the justification ratio of r_i within At is the same as the probability of r_i in Pt, referred to as a ratio $\tau = \mu(r_i) = p(r_i)$. With two interpretations per scenario, we obtain 28 experimental conditions α.

5.4 Fractions, Percentages, and Ratios

In Domain 1, the ratio τ of a proposition is an irreducible fraction varied between 1/6 and 2/3. That is, we ensured that the conclusion occurred in τ of the extensions. Besides the main objectives of the experiments, we want to show two further properties: that the scenario has limited influence on the results, and that the ratio — rather than the number of extensions — is the key factor that influences user believability ratings. For demonstrating the latter, we introduce redundant equivalent fractions γ (e.g., $1/2, 2/4, 3/6$) corresponding to the ratios τ using experimental conditions with 2,3,4, or 6 extensions. Each scenario Xi_j is associated with a fraction γ.

In Domain 2, we maintain the same fractions γ but also introduce another value, ω, representing the likelihood of the event described within the content of a proposition. For example, a proposition r_a="Joe is a Republican" becomes r_b="There is 70% chance that Joe is a Republican". We vary ω between 20% and 80% percent and the overall ratio is given by the product $\tau = \gamma * \omega$. Fractions γ and percentages ω in a scenario are associated using different combinations of both low or high ω and γ, or high ω and low γ and vice-versa.

Example 2. To obtain an argument theory based interpretation, one of our scenarios presented the user with a set of premises, and grounded defeasible rules from which arguments can be formed. We then only presented the conclusions of arguments from the preferred extensions which result from our framework. In this scenario, 6 preferred extensions existed referred to as possible worlds, and the conclusion $r_a =$"Joe is a Republican" was valid in two of these extensions[5]. The user's response to the question *Given the 6 stated possible worlds, how likely is that you would believe that "Joe is a Republican"?* then represented $u_\mu(r_a)$.

To obtain a probabilistic interpretation, we presented the set of propositions to the user as a list of hypothetical messages, which included both premises and conclusions of the above argumentation framework with no particular order. We informed the users that 2 out of the 6 messages reported that "Joe is a Republican". To determine $u_p(r_a)$, the user was asked the question *If 6 messages are released, how likely is that a message would state that "Joe is a Republican"?*

For these scenarios, $\tau = 1/3$, and $\gamma = 2/6$. The justification ratio of a proposition in At corresponds to the probability in Pt in Domain 1 such that $p(r_a) = \mu(r_a) = \gamma = \tau$. In Domain 2, r_b is used instead, with $\omega = 0.7$ in both interpretations and $\omega * p(r_a) = \omega * \mu(r_a) = \omega * \gamma = \tau$.

6 Methodology and results

We ran our experiments using Amazon Mechanical Turk[6], a web service that recruits participants to complete tasks. We recruited 420 participants for the experiment from the USA[7]. Data collection was performed with a questionnaire including four experimental conditions, such that a participant would see two different scenarios, respond to questions of both problems and interpretations. Initially participants were shown a training example for the argumentation theory to provide them with a basic understanding of argumentation. Each participant was then asked to respond to four combinations of different experimental conditions (α as described in Section 5).

- Domain 1: two questions within a scenario Xi, related to conditions Xi_0 and Xi_1 and an interpretation At (or Pt).
- Domain 2: two questions within a scenario Xj, where $i \neq j$, related to conditions Xj_0 and Xj_1, and an interpretation Pt (or At respectively).

Hence, no user would respond to an interpretation At and its corresponding interpretation Pt, and each user would see two different domains. We obtained 30 responses per condition α. In the remainder of the section, we detail they hypotheses associated with each type of problem, and describe our results.

6.1 Domain 1: Hypotheses

The aim of the first set of experiments is to understand whether people agree with the outcome of the probabilistic semantics when the propositions are purely

[5] A complete experiment example can be found at `http://tinyurl.com/mtck246`

[6] Amazon Mechanical Turk: `https://www.mturk.com/`

[7] Ethical approval for these experiments was granted by the College Ethics Review Board of the University of Aberdeen on 10/08/2016

qualitative. We study the believability rating of a proposition r_i in interpretation At as the outcome of the probabilistic semantics $u_\mu(r_i)$, and in the corresponding probabilistic interpretation Pt, $u_p(r_i)$. Our hypotheses are as follows.

H1.1: There is a correlation between the believability rating of At, $u_\mu(r_i)$, and the justification ratio of the conclusions, $\mu(r_i)$, obtained via the outcome of the probabilistic semantics.
H1.2: There is a correlation between the believability rating of Pt, $u_p(r_i)$, and the probability of the information being verified $p(r_i)$.
H1.3: The two correlations in At and Pt are similar.

We also test the following secondary hypotheses:

H1.4: The scenario does not influence the results: for any two scenarios with the same fraction γ there is no difference in the believability rating.
H1.5: The number of extensions does not influence the results: for any two scenarios with same τ but different γ there is no difference in the believability rating.

6.2 Domain 1: Results

Figure 2 presents the believability ratings $u_\mu(r_i)$ and $u_p(r_i)$ recorded for Domain 1. The horizontal axis is ordered according to the fraction γ associated with the experimental conditions. We also report the ratio τ corresponding to the fraction. For each scenario, $u_\mu(r_i)$ of At is shown besides $u_p(r_i)$ of Pt. The graph uses a divergent colour palette; the neutral rating is associated with the brightest colour, ratings below correspond to participants who consider the proposition unlikely, ratings above correspond to those who consider the conclusion likely. Moving from lower to higher γ (left to right), we observe that the darker area above the neutral bars increases for both At and Pt interpretations. Within each scenario, the neutral bar is approximately within the same range, with some exceptions. This provides some initial evidence that there is a correlation between the believability ratings and fractions γ.

A Spearman's rank-order correlation was run for each scenario Xi_j to determine the relationship between the believability ratings $u_\mu(r_i)$ in At and the justification ratios $\mu(r_i) = \gamma$, the outcome of the probabilistic semantics. This non-parametric test is used since the results are not normally distributed. The test showed a positive correlation value, rs, which was statistically significant ($rs(418) = .288, p \ll 0.001$). This provides evidence for hypothesis H1.1 — that there is a correlation between the probabilistic semantics and the user believability rating of a conclusion. A similar test determined that there is a statistically significant positive correlation between the believability ratings $u_p(r_i)$ in Pt and the probabilities $p(r_i) = \gamma$ ($rs(418) = .280, p \ll 0.001$). This validates hypothesis H1.2; i.e., there is a correlation between the believability rating of a piece of information and its probability. A comparison between the two correlations was examined using a Fisher's r-to-z transformation. The overall z-score value (based on the difference between the correlations and their variance) was observed to be $z = 0.13$ with $p = 0.448$. Here, we accept the null hypotheses that the two

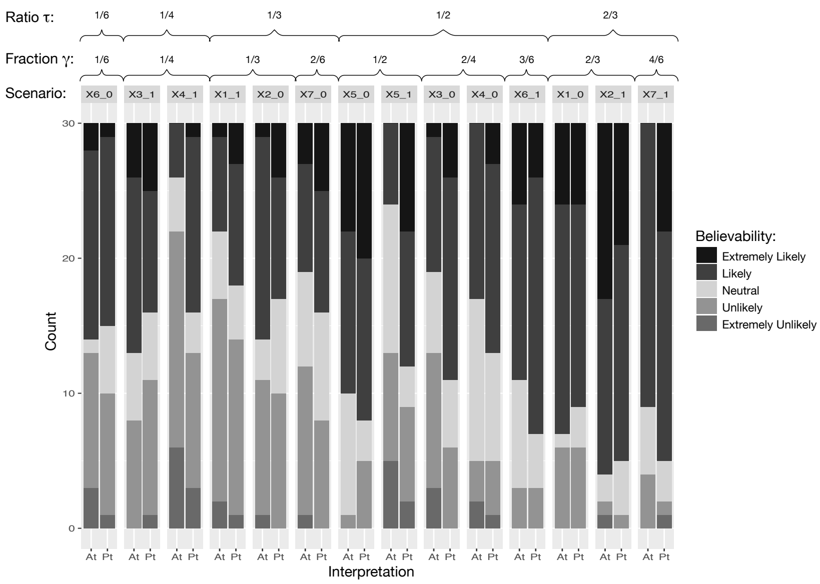

Fig. 2. Believability ratings $u_\mu(r_i)$ and $u_p(r_i)$ - Domain 1

Table 1. Mann-Whitney U tests on $u_\mu(r_i)$ vs. $u_p(r_i)$ within scenarios

Scenario	$X6\_0$	$X3\_1$	$X4\_1$	$X1\_1$	$X2\_0$	$X7\_0$	$X5\_0$
p-value	0.824	0.516	0.010*	0.265	0.888	0.247	0.744

Scenario	$X5\_1$	$X3\_0$	$X4\_0$	$X6\_1$	$X1\_0$	$X2\_1$	$X7\_1$
p-value	0.005*	0.015*	0.254	0.710	0.771	0.357	0.014*

correlations are not significantly different. This confirms hypothesis H1.3, and characterises how people interpret the outcome of the probabilistic semantics.

There are, however, some outliers that can be noticed in Figure 2. This was investigated with a post-hoc analysis using a series of Mann-Whitney U tests for each scenario Xi_j comparing $u_\mu(r_i)$ and $u_p(r_i)$. Table 1 reports only the p-values, where we consider significance at $p < 0.001$. None of the comparisons shows a significant difference, however, for the three scenarios marked with a star, the p-value tends to be low indicating the outliers.

Similar tests are used for the two secondary hypotheses. H1.4 seeks to prove that given the same fraction γ (e.g. 1/3), there is no difference between the believability rate of different scenarios associated to that fraction (e.g. $X1\_1$ vs. $X2\_0$). In Table 2 we report the p-values of comparisons between different conditions, where significant values are highlighted in bold. Hypothesis H1.4 is only partially supported: the scenario tends not to influence the results in Pt, however, in At, the hypothesis is only supported in 3 out of 5 conditions.

Hypothesis H1.5 focussed on understanding the believability ratings in experimental conditions associated with different fractions γ but same ratio τ (e.g. 1/2 for $X5\_0$ vs. 3/6 for $X6\_1$). In Table 3 we report the p-values for comparisons

Table 2. Mann-Whitney U tests on At and Pt between scenarios with similar γ

Fraction γ	X_a	X_b	$u_\mu(r_i)$ vs. $\mu(r_i)$	$u_p(r_i)$ vs. $p(r_i)$
1/4	$X3\_1$	$X4\_1$	0.403	**0.000**
1/3	$X1\_1$	$X2\_0$	0.407	0.660
1/2	$X5\_0$	$X5\_1$	0.259	**0.000**
2/4	$X3\_0$	$X4\_0$	0.669	0.208
2/3	$X1\_0$	$X2\_1$	0.147	0.056

Table 3. Mann-Whitney U tests on At and Pt between scenarios with similar τ

Ratio τ	X_a	X_b	$u_\mu(r_i)$ vs. $\mu(r_i)$	$u_p(r_i)$ vs. $p(r_i)$
1/3	$X1\_1$	$X7\_0$	0.201	0.187
1/3	$X2\_0$	$X7\_0$	0.629	0.579
1/2	$X5\_0$	$X3\_0$	0.147	**0.001**
1/2	$X5\_0$	$X4\_0$	0.068	0.006
1/2	$X5\_0$	$X6\_1$	0.417	0.526
1/2	$X5\_1$	$X3\_0$	0.932	0.370
1/2	$X5\_1$	$X4\_0$	0.677	0.016
1/2	$X5\_1$	$X6\_1$	0.574	**0.000**
1/2	$X3\_0$	$X6\_1$	0.353	0.003
1/2	$X4\_0$	$X6\_1$	0.147	0.035
2/3	$X1\_0$	$X7\_1$	0.244	0.169
2/3	$X2\_1$	$X7\_1$	0.799	**0.001**

between these conditions. H1.5 is mainly supported, with the exception of three cases in At. This provides partial evidence that it is the ratio rather than the fraction that influences the believability ratings among different conditions.

6.3 Domain 2: Hypotheses

The second problem focusses on understanding whether the outcome of the probabilistic semantics is a factor in assessing the believability of conclusions that are about event likelihood. We hypothesised that the product between the justification ratio of a conclusion and its likelihood influences people's believability ratings in the At interpretation and is comparable with the multiplication law in the probability interpretation Pt. We consider similar hypotheses as in Domain 1, with the difference that the believability ratings is now tested for correlation with the product of the fraction γ and the likelihood ω expressed within the content of a proposition ($\tau = \gamma * \omega$). Hypothesis H2.1 tests for correlation in the interpretation At where $\mu(r_j) = \tau$. Hypothesis H2.2 tests for correlation in Pt where $p(r_j) = \tau$ and H2.3 tests for similarity between the two correlations.

6.4 Domain 2: Results

Our initial tests study the correlation between the believability ratings and the fractions γ or the likelihood ω alone. Statistical tests were performed using the Spearman's rank-order correlation, and similarity is tested using the Fisher's r-to-z transformation, with significance at $p < 0.001$. We observed no correlation for fraction γ in both interpretations At ($rs(418) = .59, p = 0.228$) and Pt ($rs(418) = .26, p = 0.596$). There is, instead, a low correlation with ω in

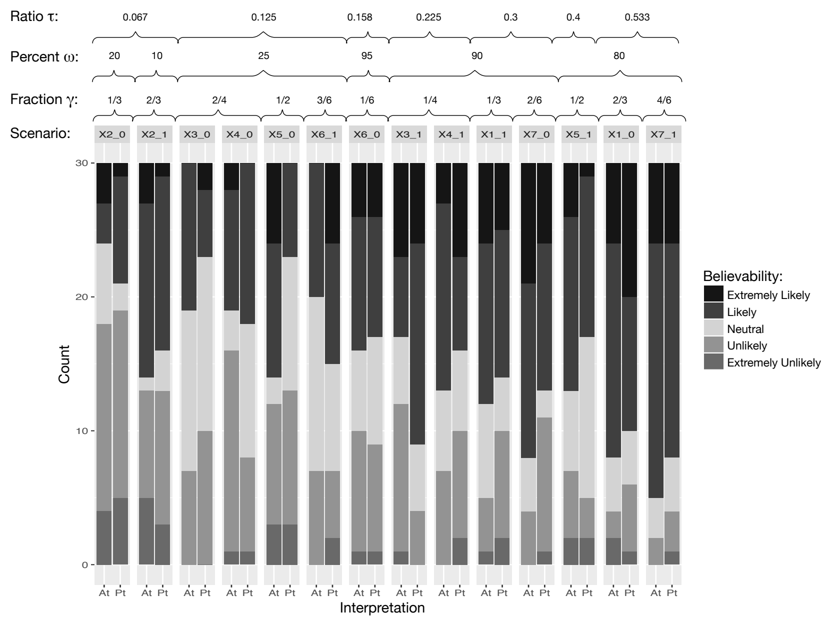

Fig. 3. Believability rating $u_\mu(r_i)$ and $u_p(r_i)$ - Domain 2

both *At* ($rs(418) = .193, p \ll 0.001$) and *Pt* ($rs(418) = .184, p \ll 0.001$) with high similarity ($z = 0.13, p = 0.448$). More interestingly, we found a correlation between the product of γ and ω reflecting the multiplication law of probability in both *At* ($rs(418) = .293, p \ll 0.001$) and *Pt* ($rs(418) = .250, p \ll 0.001$) with similar behaviour ($z = 0.67, p = 0.251$). We now focus on this last result.

In Figure 3, we present the believability rating $u_\mu(r_i)$ and $u_p(r_i)$ recorded for Domain 2. The horizontal axis is ordered according to $\tau = \gamma * \omega$. The results support hypothesis H2.1 for *At*: there is a positive correlation between the believability rating and the product of the likelihood expressed within a conclusion and the justification ratio due to the probabilistic semantics. The outcome of the probabilistic semantics is a factor required to interpret the believability ratings: the correlation with the likelihood expressed within a conclusion alone is low ($rs = .193$) and moderately improves when the product is used ($rs = .293$). A similar behaviour is observed in *Pt* supporting H2.2: there is a correlation between the believability rating and the product of the likelihood expressed within the proposition and its probability of occurring. This is stronger than the correlation with the former only ($rs = .250$ vs. $rs = .184$). Finally, H2.3 is supported as no significant difference between the two correlations values is observed.

6.5 Discussion

We have demonstrated that the outcome of the probabilistic semantics is an important factor in understanding the believability ratings of the conclusions, even in the case in which a proposition is about the likelihood of an event. The results indicate that people tend to agree with the outcome of the probabilistic semantics. Furthermore, our results confirm that the outcome of the probability semantics may be understood by people in a way similar to the understanding of probability. In the second problem, we showed that this similarity is due to a heuristic associating the product of probabilities to the believability of conclusions. Note that as discussed in Section 4, the multiplication law assumes that there is independence between the event reported by the proposition and it being inferred. We also tested for τ representing dependent events, using the law of conditional probability. The results showed no correlation with the believability ratings. Due to space constraints, we have omitted these results. The results presented here are — in a sense — preliminary. There are many aspects of this research that need further investigation. To name some, both correlation coefficients are significantly positive but show a moderate correlation between the degree of believability and the justification ratio or associated probability. This suggests that other factors need to be investigated further in the future. One of these aspects is the role of the domains used within the scenarios as we have shown that in the argumentation interpretation this has a more significant role than in the probabilistic view. From an argumentation perspective, further studies should focus on considering other semantics, such as the ranking-based semantics [3]. Further studies should also focus on understanding how people combine probabilities and on analysing human factors, for example, by considering the background of participants involved.

7 Conclusions

We investigated whether qualitative argumentation captures some notion of uncertainty by associating a degree of believability of conclusions with the number of preferred extensions. To do so, we examined whether people agree with the outcome of the probabilistic semantics. More broadly, our work can be seen to follow a strand of research similar to that of Cerutti et al. [5], aiming to study the alignment between argumentation semantics and human intuition. The novelty of our work is in that we focus on the particular role that multiple extensions play in evaluating the believability of a conclusion.

In this paper, we designed our experiments with a two-fold objective: to determine whether our claim was valid; and to investigate whether there is a similarity between probabilistic and argumentation-based reasoning. Our results showed that people tend to agree with the outcome of the probabilistic semantics and that people employ a similar heuristic in understanding both preferred extensions and probabilities. Through our experiments, we obtained some initial promising insights into the use of probability within argumentation frameworks that may guide researchers in better supporting human reasoning in their work.

Acknowledgements. This work was partially funded by a grant to the University of Aberdeen made by the UK Economic and Social Research Council; Grant reference ES/MOO1628/1.

References

1. Bailin, S., Battersby, M.: Conductive argumentation, degrees of confidence, and the communication of uncertainty. In: van Eemeren, F.H., Garssen, B. (eds.) Reflections on Theoretical Issues in Argumentation Theory, pp. 71–82. Springer International Publishing (2015)
2. Bench-Capon, T.J.M.: Persuasion in practical argument using value-based argumentation frameworks. Journal of Logic and Computation 13(3), 429–448 (2003)
3. Bonzon, E., Delobelle, J., Konieczny, S., Maudet, N.: A comparative study of ranking-based semantics for abstract argumentation. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence. pp. 914–920 (2016)
4. Cayrol, C., Lagasquie-Schiex, M.C.: Graduality in argumentation. Journal of Artificial Intelligence Research 23(1), 245–297 (2005)
5. Cerutti, F., Tintarev, N., Oren, N.: Formal arguments, preferences, and natural language interfaces to humans: an empirical evaluation. In: Proceedings of the 21st European Conference on Artificial Intelligence. pp. 207–212 (2014)
6. Croitoru, M., Vesic, S.: What can argumentation do for inconsistent ontology query answering? In: Liu, W., Subrahmanian, V.S., Wijsen, J. (eds.) Scalable Uncertainty Management, Lecture Notes in Computer Science, vol. 8078, pp. 15–29. Springer Berlin Heidelberg (2013)
7. Dunne, P.E., Hunter, A., McBurney, P., Parsons, S., Wooldridge, M.: Weighted argument systems: Basic definitions, algorithms, and complexity results. Artificial Intelligence 175(2), 457 – 486 (2011)
8. Haenni, R.: Probabilistic argumentation. Journal of Applied Logic 7(2), 155–176 (2009)
9. Heatherly, R., Kantarcioglu, M., Thuraisingham, B.: Preventing private information inference attacks on social networks. IEEE Transactions on Knowledge and Data Engineering 25(8), 1849–1862 (2013)
10. Hunter, A., Thimm, M.: Probabilistic argument graphs for argumentation lotteries. In: Computational Models of Argument, Frontiers in Artificial Intelligence and Applications, vol. 266, pp. 313–324. IOS Press (2014)
11. Li, H., Oren, N., Norman, T.J.: Probabilistic argumentation frameworks. In: Modgil, S., Oren, N., Toni, F. (eds.) Theory and Applications of Formal Argumentation, Lecture Notes in Computer Science, vol. 7132. Springer Berlin Heidelberg (2012)
12. Mayer, J.M., Schuler, R.P., Jones, Q.: Towards an understanding of social inference opportunities in social computing. In: Proceedings of the 17th ACM International Conference on Supporting Group Work. pp. 239–248 (2012)
13. Modgil, S., Prakken, H.: The ASPIC+ framework for structured argumentation: A tutorial. Argument & Computation 5(1), 31–62 (2014)
14. Prakken, H.: An abstract framework for argumentation with structured arguments. Argument & Computation 1(2), 93–124 (2010)
15. Tang, Y., Cai, K., McBurney, P., Sklar, E., Parsons, S.: Using argumentation to reason about trust and belief. Journal of Logic and Computation 22(5), 979 (2012)
16. Thimm, M.: A probabilistic semantics for abstract argumentation. In: Proceedings of the 20th European Conference on Artificial Intelligence. pp. 750–755 (2012)
17. Walton, D., Reed, C., Macagno, F.: Argumentation schemes. Cambridge University Press (2008)
18. Zenker, F.: Bayesian argumentation: The practical side of probability. In: Bayesian Argumentation, Synthese Library, vol. 362, pp. 1–11. Springer (2013)