Combining statistics and arguments to compute trust

Paul-Amaury Matt
Imperial College London, UK
paulmatt@mac.com

Maxime Morge
Université Lille 1, France
maxime.morge@lifl.fr

Francesca Toni
Imperial College London, UK
ft@imperial.ac.uk

ABSTRACT

We propose a method for constructing Dempster-Shafer belief functions modeling the trust of a given agent (the evaluator) in another (the target) by combining statistical information concerning the past behaviour of the target and arguments concerning the target's expected behaviour. These arguments are built from current and past contracts between evaluator and target. We prove that our method extends a standard computational method for trust that relies upon statistical information only. We observe experimentally that the two methods have identical predictive performance when the evaluator is highly "cautious", but our method gives a significant increase when the evaluator is not or is only moderately "cautious". Finally, we observe experimentally that target agents are more motivated to honour contracts when evaluated using our model of trust than when trust is computed on a purely statistical basis.

Categories and Subject Descriptors

H.1 [**Information Systems**]: Miscellaneous

General Terms

Theory, Experimentation

Keywords

Agent societies and Societal issues::Trust, reliability and reputation; Agreement Technologies::Argumentation

1. INTRODUCTION

The need for agents to assess whether to trust other agents emerges in many settings, e.g. in grid computing (where agents represent consumers or providers of computational power), service-oriented architectures (where agents are service providers and requesters), and e-market places (where agents are buyers and sellers of products or goods). In all these settings, agents' agreements may be represented by *contracts*, that are meant to regulate future interactions between them. Contracts may take the form of SLAs in grid computing, WS-BPML descriptions in service-oriented architectures, and deontic norms in generic multi-agent systems. In these settings, agents are vulnerable to other agents violating contracts they have jointly agreed. Models of trust can be used to mitigate this vulnerability.

Broadly speaking, trust reflects the willingness of a given agent (referred to as *evaluator*) to engage in a relationship or interaction with another agent (referred to as *target*). Trust computing is considered by the artificial intelligence community as a problem of reasoning and decision making under uncertainty. Yu and Singh [20] have defined a popular approach to trust computing using a Dempster-Shafer belief function derived from statistical data concerning the target's behaviour. In this paper, we extend Yu and Singh's approach (for local trust rating) by allowing the evaluator to take into account, in addition to the statistical data, a number of justified claims concerning the expected behaviour of the target. These claims form the basis of the evaluator's opinions and are formally represented by *arguments* in abstract argumentation [6]. We consider two classes of arguments: *forecast arguments*, in favour or against trusting the target, and *mitigation arguments*, attacking forecast arguments or other mitigation arguments. We define a method for constructing Dempster-Shafer belief functions from statistical data *and* these arguments.

We compare experimentally the predictive performance of our new method with that of the Yu & Singh's method. In the experiments, forecast arguments rely upon the existence or lack of contract clauses regulating the behaviour of the target, and mitigation arguments rely on past violations of contract clauses by the target. We compare the percentages of correct trust decisions made by the evaluator over a large number of interactions and vary both the level of "cautiousness" of the evaluator (understood as a measure of how risk averse the evaluator is) and the level of "fraudulence" of the target (understood as the frequency of contract clause violations). We observe that, independently of the level of fraudulence of the target, our method makes significantly better predictions when the evaluator is not or is only moderately cautious. This is a valuable result since a too high level of cautioness would unnecessarily prevent many possibly fruitful interactions. We also compare the "economic benefits" that the target draws from being fraudulent towards the evaluator. We observe that, independently of the level of cautiousness adopted by the evaluator, it is more beneficial for agents to honour contracts when our model for trust is used than when trust is computed on a purely statistical basis.

The paper is organised as follows. Section 2 presents background on argumentation and the classical evidence-based

Cite as: Combining statistics and arguments to compute trust, Paul-Amaury Matt, Maxime Morge and Francesca Toni, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. 209-216

model of trust introduced by Yu & Singh [20]. Section 3 motivates and gives some preliminary definitions for our model of trust, which we present in Section 4. Section 5 instantiates our model using contract-based arguments. Section 6 presents a statistical validation protocol and Section 7 highlights some experimental results. Section 8 discusses related work and we conclude in Section 9.

2. BACKGROUND

2.1 Argumentation

An *argumentation framework* [6] is a i.e. a pair (Arg, att) where Arg is a set of arguments and $att \subseteq Arg \times Arg$ is a binary relation representing attacks between arguments. For instance, we may have a framework with $Arg = \{a, b, c, d, e, f\}$ and $att = \{(a, b), (b, a), (b, c), (c, d)\ (e, c), (f, e)\}$, shown in Fig. 1 as a so-called argument graph.

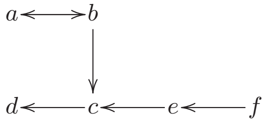

Figure 1: An example argumentation framework.

The main purpose of argumentation theory is to identify which arguments in an argumentation framework are rationally "acceptable". Several notions of acceptability have been proposed in the literature on argumentation, some providing an *intrinsic* measure of argument strength, others giving an *interaction-based* measure. Intrinsic measures are given by approaches such as [9, 14, 1, 8], whereby the acceptability of an argument depends on its internal logical structure and is independent of attacks from other arguments. On the other hand, interaction-based measures assess the strength of arguments depending exclusively on the arguments that attack them (the attackers), the attackers of these attackers (the defenders), etc. Amongst interaction-based measures, one may again distinguish between qualitative [6, 16] and quantitative [2, 4, 12] measures.

An example of qualitative measure is given by *stable extensions* [6], whereby a set of arguments X is a stable extension if and only if it is conflict-free (there is no argument in X which attacks another argument in X) and every argument that is not contained in X is attacked by (some argument in) X. In our example, $X = \{a, c, f\}$ and $X = \{b, d, f\}$ are both stable extensions.

In the remainder, we assume given, for an argumentation framework $F = (Arg, att)$, an interaction-based measure of strength $s_F : Arg \to [0, 1]$. The strength values of the arguments in the framework of Fig. 1 for some of the quantitative measures available in the literature are shown in Fig. 2. Note that we assume that, in the case of stability, the strength is assigned with respect to a chosen stable extension ($\{a, c, f\}$ first and $\{b, d, f\}$ then, in the figure), and is 1 for the arguments in this extension, and 0 otherwise. Until Section 6, we will not commit to any specific notion of strength.

measure	strength of a, b, c, d, e, f
stability [6]	1, 0, 1, 0, 0, 1 or 0, 1, 0, 1, 0, 1
graduality [2, 4]	0.618, 0.618, 0.472, 0.679, 0.5, 1
game theory [12]	0.5, 0.5, 0.417, 0.5, 0.25, 1

Figure 2: Examples of argument strength values.

2.2 Yu & Singh's trust model

Yu and Singh's evidence-based trust model [20] uses a Dempster-Shafer belief function $Bel : 2^{\Omega} \to [0, 1]$ where $\Omega = \{T, \neg T\}$ is a simple universe with T (resp. $\neg T$) representing that the *evaluator* considers the *target* to be trustworthy (resp. untrustworthy). In general, a belief functions Bel need to be defined via some evidence mass function, $m : 2^{\Omega} \to [0, 1]$, which needs to be positive, normalised and such that $m(\emptyset) = 0$. Given such m, for every subset $E \subseteq \Omega$

$$Bel(E) = \sum_{X \subseteq E} m(X)$$

In [20], the evidence mass function may be derived either from the knowledge of the evaluator's own past interactions with the target (local trust rating), or by combination of belief functions representing testimonies from other entities concerning the target (belief combination). This paper concerns only local trust rating.

Assume that the evaluator and target have a (sufficiently long) history of past interactions and that the target is capable of classifying these interactions as *poor*, *satisfying*, or *inappreciable*.[1] Denote by

- N^- the number of times the quality of the interaction was poor
- N^+ the number of times the quality of the interaction was satisfying
- $N_?$ the number of times the quality of the interaction was inappreciable
- $N = N^- + N^+ + N_?$ the total number of interactions

Then, the evidence mass function m is given by

$$m(\emptyset) = 0 \quad m(\{T\}) = \frac{N^+}{N} \quad m(\{\neg T\}) = \frac{N^-}{N} \quad m(\Omega) = \frac{N_?}{N}$$

The resulting $Bel(\{T\}) = m(\{T\})$, $Bel(\{\neg T\}) = m(\{\neg T\})$ are interpreted resp. as the trust and distrust of the evaluator in the target given the past N interactions. Note that, according to this model, trust and distrust need not sum up to 1, as we only have $m(\{T\}) + m(\{\neg T\}) + m(\Omega) = 1$ where $m(\Omega)$ can be strictly positive.

The evaluator can use the belief function Bel to decide whether to interact with the target in the following manner. Let $\rho \in [0, 1]$ be the *cautiousness* of the evaluator. The evaluator would decide to interact with the target if and only if its trust in the target exceeds its distrust by the threshold value ρ, i.e.

$$Bel(\{T\}) - Bel(\{\neg T\}) \geq \rho$$

Since $Bel(\{T\} = m(\{T\}$ and $Bel(\{\neg T\} = m(\{\neg T\}$, and since $m(\{T\}) - m(\{\neg T\}) \leq m(\{T\}) + m(\{\neg T\}) = 1 - m(\Omega)$, the larger $m(\Omega)$ and the larger ρ, the smaller the chance that the evaluator will decide to interact with the target.

[1] Yu and Singh [20] use parameters q and Q (with $q < Q$) to perform this classification: an interaction is deemed poor, satisfying or inappreciable if its quality X is such that $X \leq q$, $X \geq Q$ or $q < X < Q$ respectively. We can adopt this, or any other technique, to classify interactions.

3. STATISTICAL EVIDENCE AND ARGUMENTS

In Yu and Singh's approach to computing local trust ratings the evaluator only exploits evidence from past interactions. However, other types of evidence may be of relevance for predicting the behaviour of the target and deciding whether to engage in new interactions or not. For example, the evaluator may use witness information (word-of-mouth), social information (e.g. the social relation with the target) or prejudice (based on external signs exhibited by the target). The evaluator may wish to incorporate in its assessment these other types of evidence, in terms of arguments relative to the potential outcomes of these future interactions. Argumentation was introduced in artificial intelligence as a general-purpose paradigm for modelling reasoning with uncertain and contradictory knowledge [9]. It is thus natural to use it when reasoning about trust [15].

There is another important reason for combining statistical evidence and arguments. Yu and Singh's approach to computing local trust ratings has a theoretical predictive performance barrier. Indeed, if the target's interactions with the evaluator have constant satisfying quality frequency N^+/N and constant poor quality frequency N^-/N, the evaluator will constantly make the same trust decision under Yu and Singh's approach. If the decision were to trust (and interact), then the percentage of correct decisions would be equal to $C = N^+/N$. If the decision were to distrust (and not interact), then the percentage of correct decisions would be equal to $1-C = (N^- + N_?)/N$. Thus, the theoretical predictive performance of the model is at most $\max(C, 1-C)$.

Until Section 5, we will assume that arguments about the potential outcomes of future interactions between the evaluator and the target are given in the form of an argumentation framework F with an associated strength function s_F, as discussed in Section 2.1. In Section 5 we will see how F can be instantiated in situations where contracts are put in place to regulate interactions.

Although abstract, we will assume that arguments in F can be of one of two kinds:

- *forecast arguments*, supporting either T or $\neg T$
- *mitigation arguments*, attacking forecast arguments or other mitigation arguments.

We will denote the element of Ω *supported by* a forecast argument a as X_a.

Intuitively, forecast arguments can be seen as justified claims concerning the expected or anticipated behaviour of the target. Arguments are generally speaking not intended to be given the force of mathematical proofs [9], but serve as simple hints and clues. Therefore, the validity or strength of these arguments needs to be carefully examined by the evaluator. Mitigation arguments are here used to express the evaluator's own uncertainties concerning the validity of forecast arguments. Indeed, mitigation arguments may have the effect of reducing the strength of forecast arguments.

4. THE COMBINED TRUST MODEL

We define a new *argumentation-based belief function* in a manner similar to Yu and Singh's, but in terms of a new *argumentation-based evidence mass function* m_a combining statistical evidence and arguments as evidence. We assume as given

- statistical information N^+, N^-, $N_?$ and N,
- an argumentation framework $F = (Arg, att)$ and a measure of strength $s_F : Arg \to [0,1]$.

We will use the following notation: A is the set of all forecast arguments in F, namely $A = \{a | a \in Arg \wedge X_a \in \Omega\}$.

The argumentation-based evidence mass function m_a is defined in terms of several parameters, presented next.

- The *informational value of forecast arguments per unit of strength* parameter, V_A, indicates how much arguments should "count".

For example, setting $V_A = 0$ amounts to sanctioning that arguments have nil informational value and should be ignored. In Section 5 we will provide a definition of V_A when contracts regulating interactions are available.

- The *total amount of information* parameter I indicates how much arguments should "count" alongside statistical evidence.

We will assume that

$$I = 1 + V_A \sum_{a \in A} s_F(a)$$

namely, the statistical evidence should always count as 1, and I increases, proportionally to V_A, with the number and strength of the forecast arguments A in the argumentation framework F.

- The *indeterminacy of the evaluator* parameter, ϵ_A, gives a measure of the uncertainty of the evaluator given the past interactions with the target.

We will assume that

$$\epsilon_A = \frac{K}{K + I}$$

where K, the *epistemic risk aversion* parameter, gives a measure of the willingness of the evaluator to take risks based on its beliefs about the target, and can thus be defined as the ratio between inappreciable and poor or satisfying past interactions, namely:

$$K = \frac{N_?}{N - N_?}$$

This choice of K is motivated by the fact that known (poor or satisfying) past interactions count as (statistical) evidence. By definition of ϵ_A, the greater the epistemic risk aversion (K), the greater the indeterminacy of the evaluator. and the greater the amount of information (I), the smaller the indeterminacy of the evaluator.

The new evidence mass function is defined in terms of two priors, defined next.

Definition 1. *The* statistical prior $\hat{p} : 2^\Omega \to [0,1]$ *is defined by*

$$\hat{p}(\emptyset) = 0 \quad \hat{p}(\{T\}) = \frac{N^+}{N^+ + N^-}$$

$$\hat{p}(\Omega) = 1 \quad \hat{p}(\{\neg T\}) = \frac{N^-}{N^+ + N^-}$$

Definition 2. *The* argumentation-based prior $\hat{p}_A : 2^\Omega \to [0,1]$ *is defined by*

$$\hat{p}_A(E) = \frac{1}{I}\left[\hat{p}(E) + V_A \sum_{a \in A} s_F(a)\hat{p}(E|\{X_a\})\right]$$

where $\hat{p}(E|\{X_a\})$ is the conditional probability of E given X_a, namely $\hat{p}(E|\{X_a\}) = \dfrac{\hat{p}(E \cap \{X_a\})}{\hat{p}(\{X_a\})}$.

The argumentation-based prior extends the statistical prior by taking arguments into account. Note that, since arguments are about future behaviour and statistical data are about past behaviour, there is no double counting of information in the definition of argumentation-based prior.

It is intuitively a statistical estimator for the probabilities of the scenarios T and $\neg T$ with a bias originating from the forecast arguments. The bias introduced by each forecast argument a is all the more important as the strength of the argument $s_F(a)$ is high. Therefore, mitigation arguments have an indirect impact on the impact of forecast arguments in the computation of trust, since they may lower the strength of forecast arguments (by attacking them) or they may increase the strength of forecast arguments (by attacking other mitigation arguments that attack them).

Definition 3. *The* argumentation-based evidence mass function $m_A : 2^\Omega \to [0,1]$ *is defined by*

$$m_A(\emptyset) = 0 \quad m_A(\{T\}) = (1-\epsilon_A)\,\hat{p}_A(\{T\})$$
$$m_A(\Omega) = \epsilon_A \quad m_A(\{\neg T\}) = (1-\epsilon_A)\,\hat{p}_A(\{\neg T\})$$

As a consequence of the choices for parameters I and K, if the arguments are completely ignored (i.e. $V_A = 0$), the argumentation-based evidence mass function coincides with the classical evidence mass function of Section 2.2:

Theorem 1. *Let $V_A = 0$. Then, $m_A = m$.*

Proof. Trivially, $m_A(\emptyset) = m(\emptyset)$ (for any V_A). If $V_A = 0$, then, by our choice of I, $I = 1$, and, by definition of argumentation-based prior, $\hat{p}_A = \hat{p}$. Also, again since $I = 1$,

$$\epsilon_A = \frac{K}{K+1} = \frac{\frac{N_?}{N-N_?}}{\frac{N_?}{N-N_?}+1} = \frac{\frac{N_?}{N-N_?}}{\frac{N}{N-N_?}} = \frac{N_?}{N}.$$

$$\begin{aligned}
\text{Then,}\quad m_A(\{T\}) &= (1-\epsilon_A)\hat{p}(T)\\
&= (1-\frac{N_?}{N})\frac{N^+}{N^+ + N^-} = \frac{N^+}{N}\\
&= m(\{T\})\\
m_A(\{\neg T\}) &= (1-\epsilon_A)\hat{p}(\neg T)\\
&= (1-\frac{N_?}{N})\frac{N^-}{N^+ + N^-} = \frac{N^-}{N}\\
&= m(\{\neg T\})\\
m_A(\Omega) &= \epsilon_A = \frac{N_?}{N} = m(\Omega)
\end{aligned}$$

□

Once the new evidence mass function m_A is given, the new belief function can be directly obtained:

Definition 4. *The argumentation-based belief function $Bel_A : 2^\Omega \to [0,1]$ is defined, for every subset $E \subseteq \Omega$, by*

$$Bel_A(E) = \sum_{X \subseteq E} m_A(X)$$

Given this belief function, the evaluator still needs to decide whether to interact with the target. This can be done in the same way as in Yu & Singh's approach, given a cautiousness level $\rho \in [0,1]$, by checking

$$Bel_A(\{T\}) - Bel_A(\{\neg T\}) \geq \rho$$

Due to theorem 1, the resulting combined trust model generalises the standard evidence-based trust model (namely it agrees with it when arguments are ignored).

We conclude this section by justifying our choice of $K = N_?/(N - N_?)$. When ignoring arguments, we have as the evaluator's indeterminacy $\epsilon_A = N_?/N$ (see the proof of theorem 1). Let $\Omega' = \{T, \neg T, ?\}$, where ? represents unresolved uncertainty over the future behaviour of the target. The true probability distribution $P : \Omega' \to [0,1]$ is obviously

$$P(T) = \frac{N^+}{N} \quad P(\neg T) = \frac{N^-}{N} \quad P(?) = \frac{N_?}{N}$$

It is easy to see that $P = (1-\epsilon_A)\hat{p} + \epsilon_A P'$, where $P' : \Omega' \to [0,1]$ is defined as

$$P'(T) = 0 \quad P'(\neg T) = 0 \quad P'(?) = 1$$

Therefore, ϵ_A measures the *frequency of error* of $\hat{p}$, as this corresponds to the frequency with which Ω' follows P'.

5. ARGUMENTS FOR TRUST

In settings where trust is a critical issue, such as multi-agent systems for grid computing, service-oriented architectures and e-market places, contracts (in several formats, as discussed in Section 1) are often put in place to regulate interactions between entities and provide guarantees on the quality of interactions.

Contracts can be used to generate arguments and improve trust computing. Arguments of interest typically pertain to various "dimensions" of trust, leading evaluators to consider aspects such as *availability*, *security*, *privacy* and *reliability* of interactions. Interactions may for example amount to the provision of computational power in grid computing and the use of web-services advertised in electronic catalogues or repositories in service-oriented architectures.

For each dimension d, we consider a simple argumentation framework F_d with (some of) the following arguments:

- a forecast argument $\neg t$ supporting $\neg T$ (i.e. $X_a(\neg t) = \neg T$) on the ground that there is no guarantee in the form of a written contract clause concerning d;
- a forecast argument t supporting T (i.e. $X_a(t) = T$) on the ground that there exists a guarantee in the form of a contract clause concerning d;
- a mitigation argument v attacking t on the ground that the target has in the past "most often" violated existing contract clauses concerning d.

For each dimension d, either there exists a contract clause concerning d and t belongs to (the set of arguments in) F_d, or there does not exist such a clause and $\neg t$ belongs to F_d. If there is a contract clause, and it has been observed that, in the past, with an analogous clause the interaction with the target did not exhibit d at an acceptable level in a majority of cases, then F_d also contain v as well as an attack from v against t. As a consequence, each argumentation framework F_d contains either one or two arguments only. Using the stable extensions or game-theoretic measures of strength for these frameworks, one would assign a strength of 1 to unattacked arguments (for both measures) and a strength of 0 or 0.25 (resp.) to the argument t when it is attacked by v. Although rudimentary, these argumentation frameworks are sufficient to improve the trust decisions made using statistical data only, as we report in Section 7.

In order to assess whether to interact with the target, the evaluator needs to aggregate the decisions concerning the four dimensions (availability, security, privacy, reliability). A safe way to perform this aggregation, that we will use in Section 6, is to trust the target only if it can be trusted with respect to all dimensions simultaneously. In the remainder of this section we focus on a single dimension d.

In the context of the contract-inspired argumentation frameworks given above, we can provide a definition of the parameter V_A used in Section 4 in terms of

- $N^{+}_{\neg c}$, the total number of times the interaction quality was satisfying given that there was no contract clause
- $N^{-}_{\neg c}$, the total number of times the interaction quality was poor given that there was no contract clause
- the *conditional prior* $\hat{p}_{\neg c} : \Omega \to [0,1]$ defined by

$$\hat{p}_{\neg c}(T) = \frac{N^{+}_{\neg c}}{N^{+}_{\neg c} + N^{-}_{\neg c}} \quad \hat{p}_{\neg c}(\neg T) = \frac{N^{-}_{\neg c}}{N^{+}_{\neg c} + N^{-}_{\neg c}}$$

Definition 5. *The* informational value of forecast arguments per unit of strength *can be calculated as follows:*

$$V_A = \frac{\hat{p}(\{T\}) - \hat{p}_{\neg c}(T)}{\hat{p}_{\neg c}(T)}$$

This definition of V_A has the desirable property that, in the absence of contract clauses (and thus $\neg t$ in F_d), the argumentation-based prior for T coincides with the conditional prior for T, namely:

Theorem 2. *If $\neg t \in Arg$, then $\hat{p}_A(\{T\}) = \hat{p}_{\neg c}(T)$.*

Proof. If $\neg t \in Arg$ then by definition F_d contains no other argument (namely $Arg = \{\neg t\}$) and $s_{F_d}(\neg t) = 1$ (for any notion of strength). Thus

$$\hat{p}_A(\{T\}) \quad = \frac{1}{1+V_A}\left[\hat{p}(\{T\}) + V_A * \hat{p}(\{T\}|\{\neg T\})\right]$$

Since $\hat{p}(\{T\}|\{\neg T\}) = 0$:

$$\hat{p}_A(\{T\}) = \frac{\hat{p}(\{T\})}{1+V_A}$$

The solution V_A as given in definition 5 of the equation $\hat{p}_A(\{T\}) = \hat{p}_{\neg c}(T)$ is obtained by simple calculus. □

This property is appealing since, in the absence of contract clauses for the current interaction, the past behaviour of the target when contract clauses were absent is the only ground for an appropriate decision concerning trust.

Note that interactions take place, and thus can be evaluated, even when there is no contract given that regulates these interactions. Note also that while the statistical data is solely about the effective quality of the interactions, the arguments encompass the willingness of the target to fulfil contractual obligations.

6. EXPERIMENTAL SET-UP

To test our model against the original model of [20], we fix an evaluator and target and generate a large number of interactions with random outcomes. For each interaction,

1. we ask the evaluator to make a decision (trust or not trust) concerning the target
2. we determine the precise outcome of the interaction
3. we assign a score to the evaluator, depending on the correctness of its decision.

We assume that the evaluator decides to trust if and only if

$$Bel(\{T\}) - Bel(\{\neg T\}) \geq \rho$$

is satisfied for the model of [20], and

$$Bel_A(\{T\}) - Bel_A(\{\neg T\}) \geq \rho$$

is satisfied for our combined trust model. In both cases, ρ has a constant value over time. For Bel_A we use V_A and F as in Section 5 and s_F given by the game-theoretic measure of [12].

We will naturally be interested in the influence of ρ on the "predictive performance" of the two methods: for a fixed value of ρ and after a large number of interactions, we want to compare the percentage of correct decisions obtained using the classical evidence-based model with the percentage of correct decisions obtained with the combined trust model.

Each interaction is random, which means that its characteristics are determined randomly and also that its outcome is generated randomly from its characteristics. The basic characteristics of an interaction are the presence or absence of contract clauses guaranteeing availability, security, privacy, and reliability. The presence of each type of clause is modelled by a Boolean random variable following a Bernoulli distribution with mean 0.8. This means for instance that the probability for an interaction to have a contract clause on reliability is 80%. All four random variables are independent and identically distributed.

The randomness of the outcome of a situation with known basic characteristics is solely due to the unpredictability of the target. Each outcome is modelled as a quadruple of Beta-distributed variables denoted $\mathcal{A}$ (availability), $\mathcal{S}$ (security), $\mathcal{P}$ (privacy) and $\mathcal{R}$ (reliability). By definition of a Beta distribution, these four variables take their values inside the interval $[0,1]$. The variances of the Beta distributions are fixed to $\sigma^2 = 0.03$ but the means can take two possible values: $\mu = 0.80$ or $\mu = 0.40$. We use

- the higher mean of $\mu = 0.80$ for the distribution of a variable when there exists a contract clause guaranteeing the dimension corresponding to the variable and the target has respected the clause
- the lower mean of $\mu = 0.40$ otherwise, i.e. when there is no contract clause or when there is one but the target has ignored it.

A Beta distribution with high μ models the target's intention to produce a good interaction, and a distribution with low μ a lack of effort on the side of the target. The intention to make this effort is mainly conditioned by the presence of a contract clause. The idea is that our statistical simulation should capture the fact that targets tend to honour the contracts they sign.

In order to be realistic, the simulation also needs to take into account the possibility of "fraud", understood here as contract violation. We introduce an additional parameter θ representing the frequency with which the target is "fraudulent" (not respecting the contract). When $\theta = 0$, the target is perfectly "honest", and when $\theta = 1$, the target is never "honest".

In each situation, once the parameters of the Beta distributions have been fixed, the variables $\mathcal{A}$, $\mathcal{S}$, $\mathcal{P}$ and $\mathcal{R}$ can be sampled. The level of $\mathcal{A}$ (similarly for $\mathcal{S}$, $\mathcal{P}$ and $\mathcal{R}$) is deemed

- *sufficient* if and only if $\mathcal{A} \geq Q = 0.6$
- *insufficient* if and only if $\mathcal{A} \leq q = 0.4$
- *inappreciable* if and only if $q = 0.4 < \mathcal{A} < Q = 0.6$

Overall, the outcome is *satisfying* if and only if all four levels of availability, security, privacy and reliability are sufficient. The outcome is *disappointing* otherwise (at least one of the levels of availability, security, privacy or reliability is not sufficient). The decision by the evaluator is deemed *correct* if and only if

- it has decided to trust and interact and the outcome is satisfying, or
- it has decided not to trust and not to interact and the outcome is disappointing.

We use for the simulation a sample of 1000 situations and start registering the evaluator's score at iteration 20 only, so as to let the evaluator "learn" about the statistical behaviour of the target, in conformance with the assumption of local trust rating in [20].

7. EXPERIMENTAL RESULTS

Our first result assumes $\theta = 0$ (no fraud) and compares the percentage of correct decisions obtained with the two trust models depending on the cautiousness parameter ρ. The two curves obtained are depicted in Fig. 3.

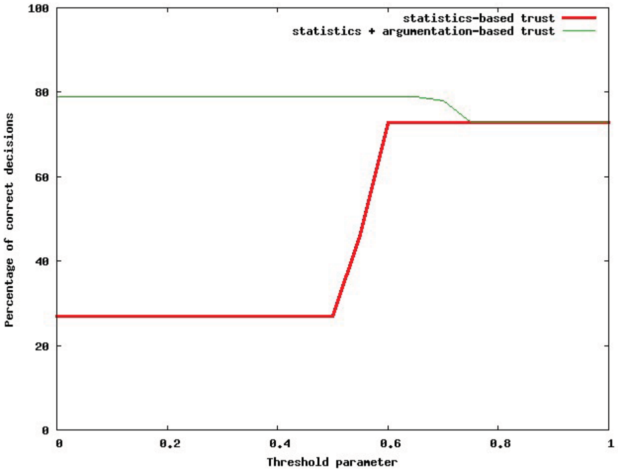

Figure 3: Performance depending on ρ and assuming $\theta = 0$ (no fraud).

The performance curve obtained for Yu & Singh's model shows a plateau at 27.04% performance for $\rho \leq 0.5$, a short performance affine transition for $\rho \in [0.5, 0.6]$ and another higher plateau at 72.86% for $\rho \geq 0.6$. Our model leads to a long performance plateau at 78.98% for $\rho \leq 0.65$, a short performance transition for $\rho \in [0.65, 0.75]$ and a slightly lower plateau also at 72.86% for $\rho \geq 0.75$. For $\theta = 0$, the performance is thus increased by 51.94% points for $\rho \in [0, 0.5]$, more than 6.12% points for $\rho \in [0.5, 0.65]$ and 0% point for $\rho \in [0.65, 1]$. The reason why no improvement is possible for large values of ρ is that such values make the evaluator so cautious that it systematically distrusts (in both models).

We have seen, in Section 3, that in Yu and Singh's model the percentage of correct decisions is theoretically limited to $\text{MaxPerf} = \max(C, 1 - C)$, where, in the context of the simulation, C is the frequency with which the outcome of a situation is satisfying. Here we have $C = 27.04\%$, $1 - C = 72.86\%$ and consequently $\text{MaxPerf} = 72.86\%$. We find that indeed Yu & Singh's trust model performance is limited by MaxPerf, but the performance of our trust model breaks that theoretical barrier and offers a level of performance that is either equal or above MaxPerf for every possible $\rho \in [0, 1]$.

Let us examine how these results change when fraud is taken into account. To compare the two models, we consider the worst-case performance measure with respect to the fraud frequency θ, for every possible value of ρ. The worst-case performance of each model is shown in Fig. 4. The worst case performance under the new model is clearly higher (irrespective of the choice of ρ), with at best an improvement of 53.47% points when $\rho \approx 0$. The improvement is null for high values of ρ (as we should have expected). The average worst-case performance improvement calculated for ρ ranging over $[0, 1]$ is of 28.21% points.

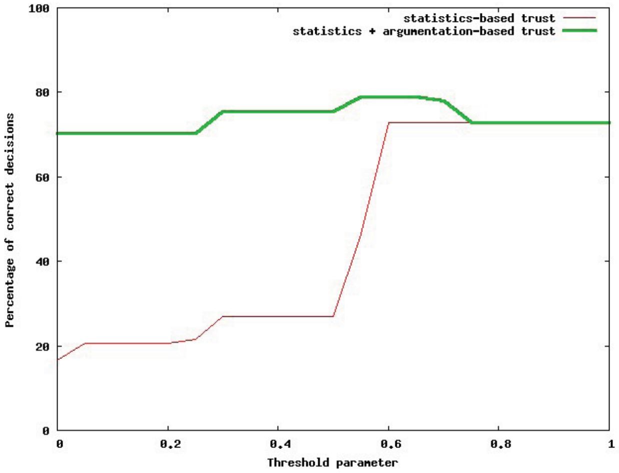

Figure 4: Worst-case performance with respect to θ (fraud frequency) depending on ρ.

Finally, we conducted an experimental study to compare the optimal level of fraudulence from the target's perspective with and without arguments being used for computing trust. For this experiment, we assume that the evaluator has a fixed level of cautiousness ρ and that the (malicious or incompetent) target has the (unlikely but theoretically possible) advantage of knowing the value of ρ. For each value of ρ, we let θ range from 0 to 1 and find the value $\theta^*(\rho)$ which maximizes the expected profits of the target. We simply assume here that for each interaction, the target makes a profit of:

- 1 unit of money if the evaluator has trusted the target and the target has not made the effort to respect the contract clauses,
- 1/2 unit of money if the evaluator has trusted the target and the target has made the effort to respect the contract clauses, and
- 0 if the evaluator has not trusted the target.

The intuition here is that when the evaluator refuses to trust the target, the target cannot charge for its "service" and thus does not make any profit. On the contrary, when the target is trusted, the evaluator will pay for the "service" provided. However, making the effort to honor all contract clauses generally induces additional costs for the target (e.g. the target may have to pay charges to third parties providing secure connections, have results post-processed to third parties to improve their reliability, etc) which reduce the net profit made by the target. Because higher profits can be made on single interactions, there exists a natural incentive for the target to fraud (infringe on contract clauses). Essentially, we want to know if argumentation reduces that incentive or not, and, if it does, to quantify the extent of this reduction. In Fig. 5, we have plotted the optimal value $\theta^*(\rho)$ of θ in function of ρ, both with and without taking arguments into account.

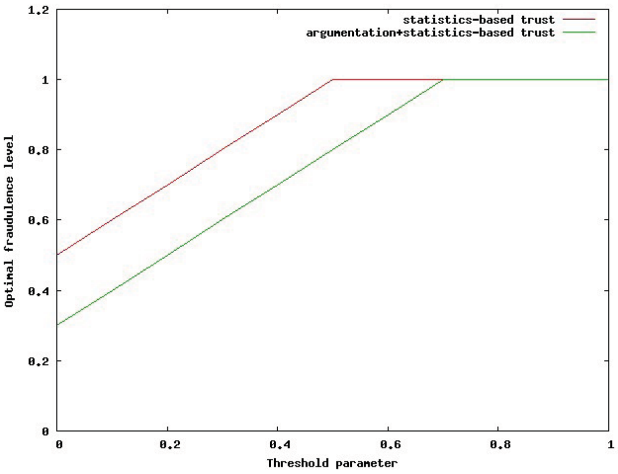

Figure 5: Optimal fraudulence level θ for the target depending on ρ.

In both cases, the optimal fraudulence level increases with the cautiousness of the evaluator. This is unsurprising as, the more cautious the evaluator, the fewer occasions the target will have to make profit, and thus, on those few occasions, a malicious target will seek to maximise its profits by violating the contract. More interestingly, we observe that, independently of the level of cautiousness adopted by the evaluator, the optimal level of fraudulence for the target is always lower when arguments are incorporated than not. The reduction of the fraudulence frequency $\theta^*(\rho)$ is:

- constant and equal to 20% in the range $\rho \in [0, 0.5]$
- reduces linearly from 20% to 0% for $\rho \in (0.5, 0.7)$
- is null for $\rho \geq 0.7$

We may therefore conclude that, although argumentation does not always reduce the level of fraud, it usually reduces it significantly and never makes its worse for evaluators. It is in this respect that our argumentation-based approach to trust contributes to a better "stability" of multi-agent systems, being understood here as the willingness of target agents to honour contracts.

8. RELATED WORK

Computing trust is a problem of reasoning under uncertainty, requiring the prediction and anticipation by the evaluator of the future behaviour of the target. Despite the acknowledged ability of argumentation to support reasoning under uncertainty (e.g. see [9]), to the best of our knowledge, only [15, 5] have considered the use of arguments for computing trust in a local trust rating setting. Dondio & Barret [5] propose a set of trust schemes, in the spirit of Walton's argument schemes, and assume a dialectical process between the evaluator and the target whereby the evaluator poses critical questions against arguments by the target concerning its trustworthiness. Their schemes and critical questions could be used as forecast and mitigation arguments, resp., in our framework. Prade [15] proposes an argumentation-based approach for trust evaluation that is bipolar (separating arguments for trust and for distrust) and qualitative (as arguments can support various degrees of trust/distrust). Our approach can be also deemed to be bipolar, as it is never the case for the specific argumentation frameworks F_d considered that an argument for trust and one for distrust occur in F_d. However, we only consider two degrees (T and $\neg T$).

Sabater and Sierra [18] classify approaches to trust as either "cognitive", based on underlying beliefs, or "game-theoretical", where trust values correspond to subjective probabilities and can be modelled by uncertainty values, Bayesian probabilities, fuzzy sets, or Dempster-Shafer belief functions. With respect to this classification, our approach can be said to be "hybrid", in that it is cognitive in that we use arguments as beliefs, and game-theoretic in that we employ a Dempster-Shafer belief function. The need for and benefits of hybrid trust models is also advocated in [19]. Our approach could be seen as a means "to extrapolate the data into the future" [19].

Castelfranchi and Falcone [3] argue against a purely game-theoretic approach to trust and in favour of a cognitive approach based upon a mental model of the evaluator, including goals and beliefs. Our construction of arguments based upon contract clauses could be deemed to be in line with the fulfilment and willingness beliefs of the model of [3], in that evaluator agents agree on contract clauses in order to fulfil their goals, and target agents agreement to contract shows their willingness to fulfil.

9. CONCLUSION

We have proposed a general method for computing trust in multi-agent systems based on combining a history of past interactions and arguments. Our method is an extension of the purely statistical method for trust by Yu and Singh [20], in that it collapses with this method when arguments are ignored. We have evaluated experimentally our combined method in comparison with the purely statistical method when arguments are built from (lack of) guarantees offered in contracts. The experimentation shows improvements resulting from the added value of arguments, even though these are very simple in our current experimental set up. Arguments are only taken into account according to their strength: the higher the strength the bigger the impact of the arguments on the computed trust values. The strength of an argument reflects its validity as a piece of evidence, in the context of all other arguments available. The choice of the set of arguments considered is thus important. A contribution of our paper is indeed the identification of a suitable set of arguments. However, note that a "bad" set of arguments may cause the argument-based trust function to be worse than the purely statistical one in the same way that a "bad" set of arguments would provide a poor knowledge base from which to draw conclusions when argumentation is used for knowledge representation.

In the future, we plan to consider more sophisticated arguments in the hope to further improve our experimental results. For example, we could consider mitigating arguments against the mitigating argument v (a past contract violation) in Section 5, on the ground that the contract violation was not responsibility of the target. Moreover, it would be interesting to use and assess our general model of trust in settings where contracts are not present, e.g. when interactions are accesses to information sources that are freely available, as in [11].

We also plan to conduct a further experimental analysis aiming at comparing the performances (in terms of correct predictions) of our combined method and the purely statistical method *over time* (rather than with respect to the cautiousness parameter ρ). Our conjecture is that the combined trust model will allow to make better predictions sooner, as it is less reliant on the size of the statistical data (learning set) than the evidence-based trust model.

In our experimentation, arguments are given rather than computed. In the future, we plan to integrate, within our experimentation environment, a system for the computation of arguments and their strength, possibly given a knowledge base from which arguments and attacks can be computed, e.g. as in assumption-based argumentation [7].

We have focused on the computation of trust by an evaluator of a target in isolation. Several approaches exist allowing agents to share information about their assessment of the trustworthiness of other agents. In this setting, agents can exchange this information in the form of arguments, e.g. as in [10]. Also, agents could exchange information on past contracts with the target, as proposed in [17].

We have used argumentation to improve the predictive capabilities of evaluators. In the future, we plan to explore how argumentation could additionally support interactions with and explanation to users in hybrid systems consisting of agents as well as humans.

Finally, our method for combining evidence in the form of statistics and arguments could be applied to other domains (beyond trust computing). As future work, it would be interesting to relate our method to other methods to combine evidence in argumentation, e.g. [13].

10. ACKNOWLEDGMENTS

This work was supported by the ARGUGRID project.

11. REFERENCES

[1] L. Amgoud and C. Cayrol. On the acceptability of arguments in preference-based argumentation. In *UAI*, 1998.

[2] P. Besnard and A. Hunter. A logic-based theory of deductive arguments. *Art. Int.*, 128(1–2):203–235, 2000.

[3] C. Castelfranchi and R. Falcone. Trust is much more than subjective probability: Mental components and sources of trust. In *HICSS*, 2000.

[4] C. Cayrol and M.-C. Lagasquie-Schiex. Graduality in argumentation. *J. of Art. Int. Res.*, 23:245–297, 2005.

[5] P. Dondio and S. Barrett. Presumptive selection of trust evidences. In *AAMAS*, 2007.

[6] P. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n-person games. *Art. Int.*, 77(2):321–257, 1995.

[7] P. Dung, R. Kowalski, and F. Toni. Assumption-based argumentation. In I. Rahwan and G. Simari, editors, *Argumentation in AI: The Book*, pages 199–218. Springer, 2009.

[8] J. Kohlas, D. Berzati, and R. Haenni. Probabilistic argumentation systems and abduction. *Ann. Math. Artif. Intell*, 34(1-3):177–195, 2002.

[9] P. Krause, S. Ambler, M. Elvang-Gøransson, and J. Fox. A logic of argumentation for reasoning under uncertainty. *Comp. Int.*, 11:113–131, 1995.

[10] G. Lenzini, N. Sahli, and H. Eertink. Agents selecting trustworthy recommendations in mobile virtual communities. In *AAMAS-TRUST*, volume 5396 of *LNCS*. Springer, 2008.

[11] E. Lorini and R. Demolombe. From binary trust to graded trust in information sources: A logical perspective. In *AAMAS-TRUST*, volume 5396 of *LNCS*. Springer, 2008.

[12] P.-A. Matt and F. Toni. A game-theoretic measure of argument strength for abstract argumentation. In *JELIA*, 2008.

[13] N. Oren, T. J. Norman, and A. D. Preece. Subjective logic and arguing with evidence. *Artif. Intell.*, 171(10-15):838–854, 2007.

[14] S. Parsons. Normative argumentation and qualitative probability. In *ECSQARU-FAPR*, 1997.

[15] H. Prade. A qualitative bipolar argumentative view of trust. In *SUM*, 2007.

[16] H. Prakken and G. Sartor. Argument-based extended logic programming with defeasible priorities. *J. of Applied Non-classical Logics*, 7:25–75, 1997.

[17] S. Reece, S. Roberts, A. Rogers, and N. R. Jennings. A multi-dimensional trust model for heterogeneous contract observations. In *AAAI*, 2007.

[18] J. Sabater and C. Sierra. Review on computational trust and reputation models. *Artif. Intell. Rev*, 24(1):33–60, 2005.

[19] E. Staab and T. Engel. Combining cognitive with computational trust reasoning. In *AAMAS-TRUST*, volume 5396 of *LNCS*. Springer, 2008.

[20] B. Yu and M. P. Singh. Distributed reputation management for electronic commerce. *Computational Intelligence*, 18(4):535–549, 2002.