A Model for Augmenting Trust Management using Argumentation

Sharmin Jalal[1], Karl N. Levitt[1], Jeff Rowe[1], Elizabeth Sklar[2,3], and Simon Parsons[2,3]

[1] Dept of Computer Science, University of California at Davis, CA, USA
sjalal@ucdavis.edu,{levitt,rowe}@cs.ucdavis.edu
[2] Dept of Computer & Information Science, Brooklyn College, City University of New York, 2900 Bedford Avenue, Brooklyn, NY 11210 USA
{sklar,parsons}@sci.brooklyn.cuny.edu
[3] Dept. of Computer Science, Graduate Center, City University of New York
365 Fifth Avenue, New York, NY 10016, USA

Abstract. Previous work has used theories of evidence to incorporate belief into trust and reputation systems. Some important questions that remain, however, are how agents might recover reputation lost in disputed transactions, and how new agents with little or no past transaction history might enter the trust network. We attempt to address these issues by extending previous work using the Dempster-Shafer theory of evidence to include formal argumentation. Reasons for past bad transaction assignment can be taken into account in new transactions and discounted by importance. New agents can participate in trust networks by forwarding evidence as arguments in a distributed reputation system. We present our preliminary model on incorporation of argumentation frameworks into trust management systems to support more complex reasoning mechanisms.

Keywords: Dempster-Shafer theory, argumentation, trust.

1 Introduction

Trust and reputation systems have gained widespread use and are increasingly important in distributed online applications. Online financial transactions, social networking and mobile *ad hoc* networks are some typical examples where trust is used to gauge the potential for successful exchange. A variety of research has modeled methods for building reputation and combining reputation into the notion of trust [9, 12, 14–17]. Defining and measuring the quality of trust, finding out domain suitable incentives to encourage participation, dealing with false identities that aim to deceive others by contaminating trust and decreasing trust transitivity are some potential research areas.

Most reputation systems assume the existence of a pre-built trust network, where some initial trust values are already in place. Very few provide a basis for deriving the initial value when past transaction records are unavailable. Another interesting area deserving attention is the recovery of reputation after a small number of anomalous bad transactions. How might a trust query distinguish between transactions with some strong reason why the outcome was bad and transactions whose outcomes were judged bad for minor reasons? In these cases, the target agent deemed bad has no way to defend itself. Again, most reputation systems consider only "witnesses" (who participate in the transaction by giving reference) or "target agents" (about whom the query pertains). But there could be other agents that are neither targets nor reference providers, yet have some relevant information that could be very helpful in making decisions about trust. Currently, these agents and their information are mostly neglected in decision making.

We assert that *argumentation* is a mechanism which gathers both complete and incomplete information from different sources and reaches a conclusion through logical reasoning. Consider the situation when a seller in a financial transaction is tagged as untrustworthy and wishes to defend himself. Argumentation allows us to logically infer the reason behind a supposed bad transaction from the propositions exchanged between the buyer and the seller agents involved.

Here, we describe our preliminary work in using argumentation to address the above areas in reputation and trust management. We extend the work of Yu and Singh [22], which proposes a distributed reputation management system using the Dempster-Shafer theory of evidence. In the following sections, we describe the background work, followed with the extensions in our model.

2 Background

In [22], the authors proposed a reputation management system which employs the Dempster-Shafer theory of evidence as the underlying computational framework. Their model applies the Dempster-Shafer belief function and Dempster's rule of combination to compute local and total belief of agents. Our model extends Yu and Singh's model [22] and resolves the scenario when an agent wants to defend himself to retrieve his past good reputation. In addition, we propose a mechanism to aggregate discrete but relevant information from trusted agents and use this in measuring belief in a specific agent. We also discuss rewards and penalties to control the flow of authentic information between agents.

In Section 2.1, we give the basic notions of the Dempster-Shafer Theory of evidence, which is the foundation for Yu and Singh's work and for our extensions. Then we describe how Dempster-Shafer theory was issued by Yu and Singh. In Sections 2.2 and 2.3, we elaborate very briefly on Yu and Singh's way of computing "local trust" from past transactions and "total trust" by combining the local trust values of neighbors. Section 3 then introduces argumentation and describes how argumentation can be used to extend the kind of reasoning possible in Yu and Singh's work. Finally, Section 4 summarizes and outlines future work.

2.1 Dempster-Shafer Theory

The seminal work on the Dempster-Shafer (DS) theory of evidence is Shafer's work "A Mathematical Theory of Evidence" [20] which is an extension of Dempster's work "Upper and Lower Probabilities induced by a Multivalued Mapping" [4]. We can say DS theory is a generalization of traditional probability theory, except that in DS theory, probabilities are assigned to sets of hypotheses instead of a single hypothesis. This property makes DS theory more expressive then simple probability theory. In DS theory, there is no relationship between believing in a hypothesis and disbelieving it. Say agent A's belief in some hypothesis is 0.8. According to DS theory, it is not necessary to assign the remaining 0.2 to be disbelief in the hypothesis, but rather it could be assigned to the set of all the possible hypotheses, indicating a lack of knowledge about them. As evidence is accumulated, the uncertainty narrows down to a subset of the entire hypothesis set [11]. Say we have two hypotheses T and $\neg T$, then $Bel(T)$ represents belief in hypothesis T, $Bel(\neg T)$ represents belief in hypothesis $\neg T$, which is disbelief in T, and $Bel(\{T, \neg T\})$ represents belief in the hypothesis T or $\neg T$, which represents a lack of belief in T or $\neg T$, or, alternatively, uncertainty about which of T and $\neg T$ is true.

Another feature of DS theory is that it does not require *a priori* knowledge, which makes it appealing in cases with no previous data.

Below we introduce the terminology upon which we base our work.

Definition 1 (Frame of Discernment). *The* Frame of Discernment Θ *is the set of exhaustive and mutually exclusive hypotheses under consideration.*

While DS theory allows for arbitrary frames of discernment, in this paper we will typically be concerned with frames of discernment that contain just a proposition and its negation $\{T, \neg T\}$.

Definition 2 (Basic Probability Assignment). *The* Basic Probability Assignment *(BPA) is a function mapping the power set of the frame of discernment to the interval between* 0 *and* 1. *The BPA of the null set is* 0 *and the summation of BPA's of all the subsets of the power set is* 1.

We can write the constraints on the basic probability assignment as follows,

$$m : 2^{\Theta} \to [0, 1]$$

where $\Theta = \{T, \neg T\}$ is the frame of discernment. We will write $\mathcal{L}$ for 2^{Θ}, and so we have:

$$m(\emptyset) = 0$$

and

$$\sum_{A \subset \mathcal{L}} m(A) = 1$$

Thus:

$$m(\{T\}) + m(\{\neg T\}) + m(\{T, \neg T\}) = 1$$

$m(A)$ is also called the basic probability number and is the measure of the belief that is committed exactly to A and does not include any belief committed to any subsets of A.

Definition 3 (Belief Function). *For a subset A, the* Belief Function $Bel(A)$ *sums the basic probability number, or total belief, of all the nonempty subsets of A which are also called the* Focal Elements *of* $Bel(A)$. *The Belief Function of A is defined as:*

$$Bel(A) = \sum_{B \subset A} m(B)$$

Thus,

$$Bel(\{T, \neg T\}) = m(\{T\}) + m(\{\neg T\}) + m(\{T, \neg T\}) = 1$$

For individual members, *Bel* and *m* are the same. Therefore, $Bel(\{T\}) = m(\{T\})$ and $Bel(\{\neg T\}) = m(\{\neg T\})$.

2.2 Local Belief from Statistical Data

[22] gives two ways to evaluate the trustworthiness of a given agent, called the target. The first is used when some other agent has sufficient previous experience with the target agent. In this case, the agent will build its local belief towards the target from the historical data. The process is as follows: after each transaction, the agent collects its user's rating about the transaction and saves the latest, say, H of them. Suppose agent A has had several past transactions with agent V, and he wants to evaluate the trust he assigns to V. Thresholds Ω and ω are defined as the *upper* and *lower* trust limits of agent A respectively. Function $f(\rho)$ returns the probability of a given value ρ, where $\rho \in \{0.0, 0.1, 0.2, \ldots, 1.0\}$ represents the quality of the services reflected in past transaction ratings for V. A's local belief towards V, according to [22] is the following:

$$\begin{aligned}
Bel(\{T\}) = \quad & m(\{T\}) = \sum_{\rho=\Omega}^{1} f(\rho) \\
Bel(\{\neg T\}) = \quad & m(\{\neg T\}) = \sum_{0}^{\rho=\omega} f(\rho) \\
Bel(\{T, \neg T\}) = \, & m(\{T, \neg T\}) = \sum_{\rho=\omega}^{\rho=\Omega} f(\rho)
\end{aligned}$$

2.3 Combining Beliefs of the Witnesses

The second approach to evaluating trustworthiness in [22] is collecting local belief from the witnesses. Suppose that A does not have many past transactions with the target V. In this model, A will ask for references from its trusted neighbors. If they have had enough transactions with V, they will already have computed their local trust and can pass that value to A. If, however, they also lack sufficient transaction history, they will pass a reference to another of their trusted agents in turn. The referenced agent then supplies its local trust about V or passes along yet another reference. The authors define a *depthLimit* as the maximum length of the referral chain. Let α be the focal element of belief function *Bel* over $\mathcal{L}$. *Bel*1 and *Bel*2 are two belief functions over $\mathcal{L}$ based on different evidence. $m1$ and $m2$ are the BPA's of *Bel*1 and *Bel*2, respectively. According to *Dempster's rule of combination*, $m = m1(\alpha) \oplus m2(\alpha)$ will be the new combined BPA over α, which is the sum of the form $m1(X)m2(Y)$, where X and Y range over all subsets whose intersection is α. Therefore,

$$m(\emptyset) = 0$$

and

$$m(\alpha) = \frac{\sum_{X_i \cap Y_{j=\alpha}} m1(X_i)m2(Y_j)}{1 - \sum_{X_i \cap Y_{j=\emptyset}} m1(X_i)m2(Y_j)}$$

Here, $\{X_1, X_2, \ldots X_n\}$ are the focal elements of $m1$, and $\{Y_1, Y_2 \ldots Y_m\}$ are the focal elements of $m2$. And,

$$\sum_{X_i \cap Y_{j=\emptyset}} m1(X_i)m2(Y_j) < 1$$

is also called *conflict*. This indicates the conflict between two distinct bodies of evidence.

In the model, τ and π are defined as the functions that return the local belief and total beliefs of an agent, respectively. Therefore, in the presence of the witnesses $\Delta = \{w_1, w_2, \ldots w_n\}$, agent A will update its total belief over V, considering all of the local beliefs from its witnesses.

$$\pi_A = \tau_{w_1} \oplus \tau_{w_2} \oplus \ldots \oplus \tau_{w_n}$$

As threshold for trustworthiness is then defined. Agent A will trust agent V if,

I. $\tau_A(\{T_V\}) - \tau_A(\{\neg T_V\}) \geq$ trust threshold, in the case when agent A constructs its local belief from its own historical data.
II. $\pi_A(\{T_V\}) - \pi_A(\{\neg T_V\}) \geq$ trust threshold, when agent A constructs its total belief, combining the local belief of its witnesses.

Having described the approach suggested by Yu and Singh, we will go on to describe how argumentation can be used to extend the model.

3 Argumentation to Compute Trust

In the following sections, we will first describe the basic ideas of argumentation frameworks and the acceptability semantics. In later sections, we will describe our model in different scenarios.

3.1 Argumentation Background

In this subsection, we briefly describe some key elements of *argumentation*. We follow Dung's notions of argumentation [5], where an argumentation framework is an abstract entity whose role is determined by its relation to other arguments.

Definition 4 (Argumentation Framework). *An* argumentation framework *is a pair:*

$$AF = \langle AR, R \rangle$$

where AR is the set of arguments and R is the binary attack relation between arguments. That is, $R \subseteq AR \times AR$.

For two arguments A and B, we say A attacks B if $(A, B) \in R$.

To illustrate further the notion of argumentation, we are considering a particular argumentation system stated in [2] that handles inconsistency in the knowledge base. According to [2], arguments are built from a propositional knowledge base Σ that could be inconsistent. $\vdash$ stands for classical inference and $\equiv$ stands for logical equivalence.

Definition 5 (Argument). *[2] An argument is a pair (H, h), where $H \subseteq \Sigma$ such that*

$$H \vdash h$$

H is assumed to be consistent and minimal (for set inclusion). H is called the support, and h is the conclusion of the argument.

To illustrate the attack relation a little more, [7] defined two relations, *Rebut* and *Undercut*, which are as follows:

Definition 6 (Rebut). *Let $(H1, h1)$ and $(H2, h2)$ be two arguments.* $(H1, h1)$ rebuts $(H2, h2)$ *iff* $h1 \equiv \neg h2$.

Definition 7 (Undercut). *Let $(H1, h1)$ and $(H2, h2)$ be two arguments.* $(H1, h1)$ undercuts $(H2, h2)$ *iff $\exists h \in H2$ such that $h \equiv \neg h1$.*

Though the definition of attack in [2] includes the notion of rebut, we do not use rebut here because it has been shown to have some unfortunate consequences for argumentation systems using propositional logic [1].

Definition 8 (Conflict Free). *We say, a set S is* conflict-free *if $\forall A \in S$, $\nexists B \in S$ such that $(B, A) \in R$*

Definition 9 (Acceptable). *An argument A is* acceptable *with respect to a set S iff $\forall B \in AR$, if $(B, A) \in R$, then $\exists C \in S$ such that $(C, B) \in R$.*

That is, an argument is acceptable to a rational agent, iff he can defend that argument from his own knowledge base.

Definition 10 (Admissable). *Consider S as a conflict-free set of arguments in the framework $\langle AR, Attacks\rangle$. S is* admissible *iff each argument in that set is acceptable with respect to set S.*

Definition 11 (Preferred Extension). *A* preferred extension *is the maximal (with respect to set inclusion) admissible set of the argumentation framework AF.*

Example 1. Let $AF = \langle\{A, B, C\}, \{(B, A)(C, B)\}\rangle$. Clearly, here the preferred extension $E = \{A, C\}$.

Example 2. Let $AF = \langle\{A, B\}, \{(A, B), (B, A)\}\rangle$. There are two preferred extensions, $\{A\}$ and $\{B\}$.

Example 3. Let $AF = \langle\{A, A\}, \{(A, A)\}\rangle$. Here the preferred extension is the empty set.

Definition 12 (Stable Extension). *A conflict-free set of arguments S will be a* stable extension *(SE) iff $S = \{A | \forall B \notin S$ will be attacked by $S\}$.*

In examples 1 and 2 above, the preferred extensions are also stable extensions. But in example 3, the empty set is not a stable extension.

The preferred and stable extensions are considered to both be *credulous* — they consider an argument to be acceptable when a more skeptical approach might not. Argumentation also has more skeptical notions of an extension which we will introduce below.

Definition 13 (Characteristic Function). *Dung defined a monotonic* characteristic function F_{AF} *that returns the* acceptable sets *for each input set.*

That is,

$$F_{AF} : 2^{AR} \rightarrow 2^{AR}$$

$$F_{AF}(S) : \{A | A \text{ is acceptable with respect to } S\}$$

Dung also showed that if the argumentation framework is finitary which is, for each argument, there are a finite number of arguments that attack it, then F_{AF} is continuous and its least fixed point can be can be found by iteratively applying it to the empty set.

Definition 14 (Complete Extension). *An admissible set S is a* Complete Extension *iff all arguments defended by S are also in S.*

There could be more than one complete extension each corresponding to a particular viewpoint.

Definition 15 (Grounded Extension). *A conflict-free set of arguments S is the* Grounded Extension *if it is the minimal (with respect to set inclusion) complete extension.*

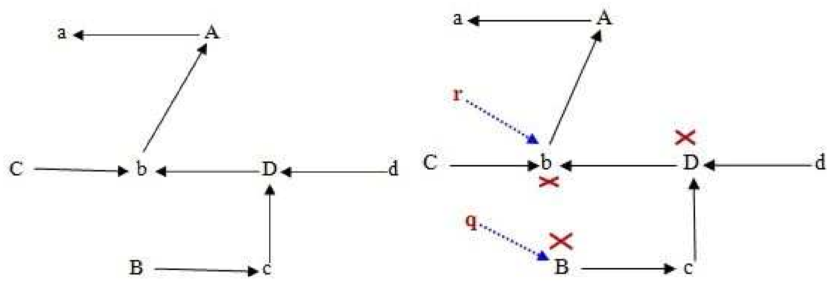
defeated as well as all those arguments that are supported directly or indirectly by these un-attacked arguments. A grounded extension is also the *Least Fixed Point* of F_{AF}. In example 1, $\{A, C\}$ is also the grounded extension, but in example 2, the grounded extension is empty. In other words, we can say that a skeptical reasoner will conclude nothing if the grounded extension is empty.

3.2 First Scenario: Target Has No Historical Data

In this scenario, we consider the situation when an agent needs to transact with another one with whom he has no previous experience and no referral. Suppose, a buyer X_i has to buy a product from seller Y_j, Neither X_i nor any of his neighbors have any previous transactions with Y_j. How is X_i going to decide if he will trust Y_j or not? Consider the conversation between X_i and Y_j to be as follows:

$$X_i's \text{ claims: } \{A, B, C, D\}$$
$$Y_j's \text{ claims: } \{a, b, c, d\}$$

From Figure 1(a), we can see that Y_j's claim a is attacked by X_i's A. Y_j backed up his claim with b, which attacks A. b is again attacked by X_i's claims C and D. In the same way, D is attacked by d and c, and B attacks c. At this point, X_i will build an argumentation framework AF out of the conversation and compute the stable extensions. Here we are assuming a meaningful argument should be well-founded and coherent. As we



(a) graphical representation of dialogue

(b) argumentation framework, AF, after providing evidence

Fig. 1. Conversation between X_i and Y_j

are also assuming all agents are skeptical reasoners, they will decide nothing unless a winning extension is found.

X_i's AF and the generated stable extensions will look like the following (we follow Dung's notion of abstract argumentation here):

$$\begin{aligned} F &= \langle Arg, Att\rangle \\ &= \langle \{A, B, C, D, a, b, c, d\}, (A, a), (b, A), (C, b), (D, b), (d, D), (c, D), (B, c)\}\rangle \end{aligned}$$

where $SE_{X_i} = \{A, B, C, D\}$ and $SE_{Y_j} = \{a, b, c, d\}$.

As our agents are skeptical reasoners, they tend to follow the grounded extension as its conclusions are not controversial. Figure 1(a) shows, X_i has two unattacked claims B and C and Y_j has one unattacked claim d. At this point, we can say X_i has two arguments that no one can attack but can we say that X_i has two arguments that no one can disprove? No. If we follow this process, anyone in Y_j's place can provide as many arguments as he can for the sake of winning. Unfortunately, this could happen both ways around and could go on and on, which will destroy the well-founded structure of the framework.

Instead of computing the conventional grounded extension (GE), we propose an *extendedGE*, which is limited to considering *evidence* as the starting point. In our model, *evidence* is the vital element to win an argument. Our opinion is that, if someone is saying something true he should be able to support his claim with evidence. Here by "evidence", we are indicating statements about the ground truth of the domain which are non-conflicting if the domain is consistent. At this point, neither X_i nor Y_j has provided any evidence. This forces them to supply evidence to fortify their claims. In Figure 1(b), we see that, X_i supports his claim C with evidence r. As evidence provides the non-conflicting ground truth, the attacked arguments will automatically be eliminated. Therefore, r eliminates b. Likewise, Y_j supports his claim c with evidence q, which eliminates B. The agent must iteratively return arguments which are themselves evidence or have evidence as a supporting argument. These will eliminate the arguments that these arguments attack. This process follows until the agents reach a conclusion. After elimination, the new stable extensions will look like the following:

$$SE_{X_i} = \{A, C, \mathbf{r}\}$$
$$SE_{Y_j} = \{a, c, d, \mathbf{q}\}$$

At this point, both of the stable extensions have exactly one piece of evidence. In our model, the evaluator will break this tie by considering the depth of the supporting evidence. We extend the idea of $depthLimit_R$ from Yu and Singh [22] and propose $depthLimit_E$ which denotes the number of hops the evidence is away from the claim it is supporting. In the example above, r is two hops away from the initial claim A, and q is six hops away from its initial claim a. Intuitively, evidence is more relevant if $depthLimit_E$ is short, and evidence becomes more irrelevant as $depthLimit_E$ increases. This makes SE_{X_i} the winner. Therefore, Y_j fails to defend his claims, and X_i will rate Y_j from its lower trust limit range which will be used later to compute the BPA of Y_j, and afterwards belief in Y_j.

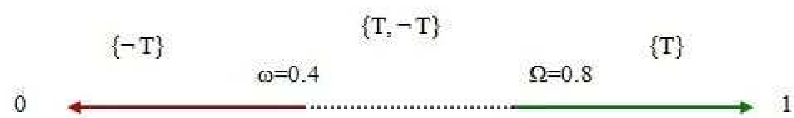


Fig. 2. Trust Scale of agent X_i

Consider the following example: X_i's upper trust limit is $\Omega = 0.8$ and lower trust limit is $\omega = 0.4$. Therefore, all the transactions with $\rho = [0.8, 1.0]$ count for $\{T\}$ and $\rho = [0, 0.4]$ count for $\{\neg T\}$. The rest count for $\{T, \neg T\}$. The scenario we present is a special case where X_i has no previous data about Y_j. X_i will select a value from its lower trust limit range $[0, 0.4]$, depending on how badly Y_j failed to defend himself as a prior rating for Y_j. Let X_i select 0.2 as the initial rating for Y_j and X_i's probability of making a good decision as 0.8. A potential way of measuring initial local belief Bel in Y_j could be:

$$Bel(Y_j) = \frac{\text{Good decisions taken by } X_i}{\text{Total decisions taken}} \times \text{ Prior rating for } Y_i = 0.8 \times 0.2 = 0.16$$

Though X_i's probability of making a good decision is high, the result is low due to Y_j's poor rating. If this value is below the risk threshold, then X_i will not engage in any communication or transactions with Y_j.

The idea of $depthLimit_E$ to count the number of hops across pieces of evidence could later be used in risk analysis. As we said before, evidence is more relevant when it supports claims closer to the primary claim. Hence, we can say:

$$depthLimit_E \propto risk$$

Some researchers propose a semantics (the ideal semantics) that is less skeptical then grounded extension but more skeptical than preferred extension [6]. In our model, we can control the skepticism by taking $depthLimit_E$ into account. Intuitively:

$$depthLimit_E \propto \frac{1}{skepticism}$$

We assume that each agent has its distinct risk threshold which solely depends on the current state of that agent. A higher risk threshold indicates the agent is capable of taking more risk. Therefore we can say that an agent with a high risk threshold can choose to consider evidence with larger $depthLimit_E$. Thus $depthLimit_E$ could be a potential factor to consider in analyzing trust sensitivity. Note that we reserve discussion of risk analysis and trust sensitivity for future work.

3.3 Second Scenario: Target has Transaction History

In this section, the seller is known to the buyer. As the buyer has had previous transactions with the seller, it will build its local trust from the previous trust rating using Dempster-Shafer theory. Consider the following: X_i has had six previous transactions with Y_j. After the last transaction, Y_j's ratings are, say, $\{0.2, 0.6, 0.9, 0.7, 0.3, 0.2\}$. Let $x \in \{0.2, 0.6, 0.9, 0.7, 0.3, 0.2\}$. According to Dempster-Shafer theory, Y_j's BPA will be:

$$\begin{aligned} m(\{T\}) &= \sum_{\Omega=0.8}^{1} f(x) = 1/6 = 0.167 \\ m(\{\neg T\}) &= \sum_{0}^{\omega=0.4} f(x) = 1/6 \times 3 = 0.5 \\ m(\{T, \neg T\}) &= \sum_{\omega=0.4}^{\Omega=0.8} f(x) = 1/6 \times 2 = 0.333 \end{aligned}$$

Therefore, the belief values for Y_j would be:

$$\begin{aligned} Bel(\{T\}) &= 0.167 \\ Bel(\{\neg T\}) &= 0.5 \\ Bel(\{T, \neg T\}) &= 0.333 \end{aligned}$$

As we can see, $Bel(\{T\}) - Bel(\{\neg T\})$ is negative (-0.333), which is obviously a lot less then the trust threshold. Thus the buyer will not engage in any transactions with the seller.

In the equation used for deciding "to trust" or "not to trust":

$$Bel(\{T\}) - Bel(\{\neg T\}) \geq trust\ threshold$$

if the difference is large (the seller is either highly trusted or highly distrusted), then it will follow the same process. But if the difference is small, which is, a big number of transactions fall under an uncertain state, then the buyer will follow the process of the first scenario, outlined in Section 3.2, to see if the current transaction can limit the uncertain state.

3.4 Third Scenario: Combining Trust in a Prebuilt Trust Network

In most practical cases, the evaluator or buyer does not have enough transactions or has no transactions at all with the desired seller. Here, buyers often look for *referrals* to the learn something about the seller. The situation where no referrals are available was described in Section 3.2. Now we are going to describe the case of combining referrals. In our model, the buyer or evaluator sends out the query to its trusted neighbors asking for testimonies about the seller. If the neighbors have past experience and have built a local belief structure (in the way described in Section 3.2), then they pass their belief value(s)[4] to the buyer. In cases where the seller is also unknown to the neighbor, the neighbor may pass a referral on to a potential agent who may have past experience with the seller. This process follows until the evaluator gets the desired testimony (or there are no more agents left to query). As mentioned earlier, in [22], the authors present $depthLimit_R$, which denotes the length of the referral chain. We introduce some additional constraints here. If $depthLimit_R$ falls outside of a given range, then the seller will be treated as a newcomer with no referral history; and the scenario described in Section 3.2 will be followed. This range will be set by risk analysis, which is a topic we reserve for future work.

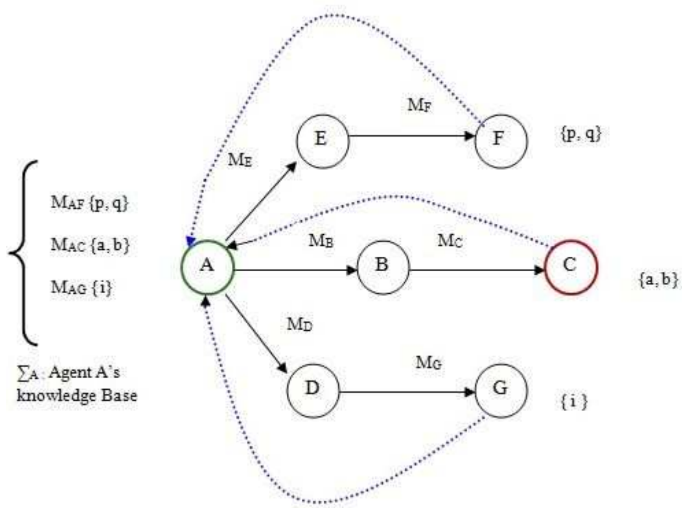


Fig. 3. Local trust propagation in pre-built trust network

[4] Multiple belief values may exist, for example, where beliefs are contextualized and a vector associates individual beliefs with a set of contexts. Here, we abstract the notion of belief into a single value and reserve discussion of belief as a complex data structure for future work.

Consider the graph in Figure 3. Our buyer, A, sends out a query about seller C to its trusted neighbors B, D and E. Among them, only B has previous experience with C and has hence built a local belief structure about C. This local belief will be passed on as a testimony to A. The transaction will be between A and C; and C's dialogues, along with its testimony, will be passed to A. At the same time, D and E will pass the query to G and F, respectively. It is a very common scenario in practical cases that G and F have no information about C, but they do have experience about the product he is selling—which is crucial in making a decision, but was not explicitly requested in the query. These claims, along with the testimonies, will be passed to A in a similar fashion.

Local belief values will be merged using the method proposed in [21]. We use the *concatenation* and *aggregation* operators proposed in [13], and subsequently used by [21], to merge the trust values in the graph. The concatenation operator is used to merge trust within the same referral chain. On the other hand, the aggregation operator is used to combine the trust values on the same topic that come from different sources (agents). In our example, consider that A's local belief towards its trusted neighbors B, D and E are M_B, M_D and M_E, respectively. Again, B holds M_C, its local belief structure concerning C, E holds M_F, its beliefs in F, and D holds M_G, its beliefs in G. Here, $M_B = Bel_B$. This belief function has three parts: "Belief" in B, "Disbelief" in B. and "Uncertainty" about B. Separately, $Bel(\{T\}) = m(\{T\})$, $Bel(\{\neg T\}) = m(\{\neg T\})$ and $Bel(\{T, \neg T\}) = m(\{T, \neg T\})$, which are the summation of the probabilities of "Good transactions", "Bad transactions" and "Uncertainty", respectively (as discussed above). For simplicity and similarity, we will follow the notions used in [21]. Let,

$$
\begin{aligned}
m(\{T\}) &= P_B \\
m(\{\neg T\}) &= N_B \\
m(\{T, \neg T\}) &= U_B
\end{aligned}
$$

Therefore, following [21], we construct A's primary beliefs about C as follows:

$$
\begin{aligned}
M_{AC} &= M_B \otimes M_C \\
P_{AC} &= P_B \times P_C \\
N_{AC} &= P_B \times N_C \\
U_{AC} &= 1 - (P_B \times P_C) - (P_B \times N_C)
\end{aligned}
$$

Here, $\otimes$ is the concatenation operator, which is just Dempster's rule from before. At this point, we can say that A has M_{AC} primary belief in C's claim $\{a, b\}$. M_{AF} and M_{AG} will be constructed in a similar way. We mention A's "primary belief" in C because A still has to assimilate all of the information he gets from G and F to come up with his final belief.

Next, all these dialogues from C, F and G will be put in an argumentation framework, along with A's knowledge similar to the scenario described in Section 3.2, except that the pivotal point will be the "combined local belief" in those claims. That means A will consider the following in constructing the argumentation framework:

$$
M_{AC}\{a, b\} \cup M_{AG}\{i\} \cup M_{AF}\{p, q\} \cup D_A
$$

Here, D_A is A's knowledge about the domain. Following our earlier assumption, high-valued claims will be prioritized over low-valued claims, while defeating each other. If there is a tie (same combined trust), then the scenario in which there is no prior history (Section 3.2) will be followed again, and this time "evidence" will be used as the tie-breaker.

3.5 Discussion

This section highlights three issues that have not been specifically addressed above, but need to be considered when using argumentation to compute trust. These issues are: the Fake Profile problem, the Trust Transitivity problem, and the Incentive problem. Each is discussed briefly, below.

The *Fake Profile* problem is a major issue in reputation systems. Membership in most social networking and business rating sites such as Yelp[5], for example, is free. As a result, there is very little to stop people creating many different profiles with which they boost or downgrade the reputation of an entity. These fake profiles have a very bad impact on cooperation or even initiation of a transaction. This impacts how newcomers will be treated [8]. In our model, every agent has to defend his claims with evidence. No matter how many profiles that agent has received or how many good referrals were collected, in the end, he needs to hold evidence. This requirement suppresses fake profiles to a great extent. Moreover, as shown in [18], with enough exchanges of arguments, it is not possible for one agent to deceive another indefinitely— eventually their knowledge bases converge.

In belief systems, *Trust Transitivity* is another major issue. It is possible that what the evaluator decides is most heavily influenced by its witnesses' beliefs. In this case, making decisions that depend upon witnesses' local beliefs is prone to deception. There are several proposals in the literature addressing trust transitivity [3, 10, 14, 19, 22]. In our model, since contributing agents are invisible to each other (e.g., in Figure 3, C, F and G are invisible to each other), a malicious agent does not gain any advantage by deceiving a trusted node, as he does not necessarily know who opposes his claim. This leaves him with no choice but to deceive large numbers of agents, possibly all of an evaluator's trusted nodes! And, in the end, the deceiver is required to show evidence; trust transitiveness does not help him much here.

There are cases when trusted agents exist but have little *incentive* to contribute information to third-party transactions. To encourage them to participate, we propose $rewardVal$ and $penaltyVal$, respectively. The latter, $penaltyVal$ will decrease the agent's rating and hence belief in him. Similarly, the former, $rewardVal$, will increase this rating. This will incentivize agents to contribute and will penalize malicious agents for infusing unauthentic information. The risk threshold of the agent can thus be compared to the $penaltyVal$ and $rewardVal$ to optimize decision making. If these values are made public, then it is possible to guess an agent's current state by analyzing these values. Moreover, if an agent is willing to deceive and can afford the $penaltyVal$ (i.e., $penaltyVal < riskThreshold$), then he may take the risk of deceiving the evaluator agent. We will discuss these values more in our future work.

[5] http://www.yelp.com

4 Summary and Future Work

In our model, we have addressed some problems in current trust management and reputation systems by incorporating *evidence* into an argumentation framework, and then integrating it into multiple trust management scenarios. In future work, our plan is to refine the theory and focus more on risk analysis. In particular, we are considering adding the concept of *utility* to our trust management models in order to capture the differential importance of evidence to different agents. This might be used to perform a risk analysis to judge the effects of making incorrect trust-based judgments. We also intend to investigate foundations and formulations for assigning trust thresholds and choosing ratings to measure BPA which will make our model more precise. Later, we plan to implement it in a more practical environment, such as a recommendation system for online social network applications.

5 Acknowledgments

Research was partially funded by the National Science Foundation, under grant CNS 1117761 and Army Research Laboratory and Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory, the National Science Foundation, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

1. L. Amgoud and P. Besnard. Bridging the gap between abstract argumentation systems and logic. In *Proceedings of the Third International Conference on Scaleable Uncertainty Management (SUM)*, Washington, DC, USA, September 2009.
2. L. Amgoud and C. Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence*, 34:197–216, 2002.
3. S. Buchegger and J. Y. L. Boudec. A Robust Reputation System for P2P and Mobile Adhoc Networks. In *Proceedings of the Second Workshop on the Economics of Peer-to-Peer Systems*, 2004.
4. A. P. Dempster. Upper and Lower Probabilities induced by a Multivalued Mapping. *Annals of Mathematical Statistics*, 38(2):325–339, 1967.
5. P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Journal of Artificial Intelligence*, 77(2):321–358, 1995.
6. P. M. Dung, P. Mancarella, and F. Toni. A dialectic procedure for sceptical, assumption-based argumentation. In *Proceedings of the 1st International Conference on Computational Models of Arguments (COMMA)*, pages 145–156, Liverpool, UK, 2006. IOS Press.
7. M. Elvang-Gøransson, P. Krause, and J. Fox. Dialectic reasoning with inconsistent information. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, 1993.
8. E. J. Friedman and P. Resnick. The social cost of cheap pseudonyms. *Journal of Economics & Management Strategy*, 10(2):173–199, 2001.

9. J. Golbeck. Combining Provenance with Trust in Social Networks for Semantic Web Content Filtering. In *International Provenance and Annotation Workshop (IPAW)*, Chicago, IL, USA, May 2006.
10. J. Golbeck, B. Parsia, and J. Hendler. Trust networks on the semantic web. In *Proceedings of Cooperative Intelligent Agents*, August 2003.
11. J. Gordon and E. H. Shortliffe. *The Dempster-Shafer Theory of Evidence*, pages 272–292. Addison-Wesley, Reading, MA, USA, 1984.
12. C.-W. Hang, Y. Wang, and M. P. Singh. An adaptive probabilistic trust model and its evaluation. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2008.
13. A. Jøsang. A subjective metric of authentication. In *Fifth European Symposium on Research in Computer Security*, Louvain-la-Neuve, Belgium, September 1998.
14. S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The EigenTrust Algorithm for Reputation Management in P2P Networks. In *Proceedings of the Twelfth International World Wide Web Conference (WWW)*, Budapest, Hungary, May 2003. ACM.
15. Y. Katz and J. Golbeck. Social network-based trust in prioritized default logic. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, 2006.
16. U. Kuter and J. Golbeck. SUNNY: a new algorithm for trust inference in social networks using probabilistic confidence models. In *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI)*, 2007.
17. P.-A. Matt, M. Morge, and F. Toni. Combining statistics and arguments to compute trust. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2010.
18. S. Parsons and E. Sklar. How Agents Alter Their Beliefs After an Argumentation-Based Dialogue. In *Proceedings of the Workshop on Argumentation in Multiagent Systems (ArgMAS) at Autonomous Agents and MultiAgent Systems (AAMAS)*, 2005.
19. T. Riggs and R. Wilensky. An algorithm for automated rating of reviewers. In *Proceedings of the First ACM/IEEE-CS joint conference on Digital libraries*, 2001.
20. G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.
21. Y. Wang and M. P. Singh. Trust representation and aggregation in a distributed agent system. In *The Twenty First National Conference on Artificial Intelligence (AAAI)*, 2006.
22. B. Yu and M. P. Singh. Distributed reputation management for electronic commerce. *Computational Intelligence*, 18(4):535–549, November 2002.