Formal Trust Model for Multiagent Systems*

Yonghong Wang and **Munindar P. Singh**
Department of Computer Science
North Carolina State University
Raleigh, NC 27695-8206, USA

Abstract

Trust should be substantially based on evidence. Further, a key challenge for multiagent systems is how to determine trust based on reports from multiple sources, who might themselves be trusted to varying degrees. Hence an ability to combine evidence-based trust reports in a manner that discounts for imperfect trust in the reporting agents is crucial for multiagent systems.

This paper understands trust in terms of belief and certainty: A's trust in B is reflected in the strength of A's belief that B is trustworthy. This paper formulates certainty in terms of evidence based on a statistical measure defined over a probability distribution of the probability of positive outcomes. This novel definition supports important mathematical properties, including (1) certainty increases as conflict increases provided the amount of evidence is unchanged, and (2) certainty increases as the amount of evidence increases provided conflict is unchanged. Moreover, despite a more subtle definition than previous approaches, this paper (3) establishes a bijection between evidence and trust spaces, enabling robust combination of trust reports and (4) provides an efficient algorithm for computing this bijection.

1 Introduction

In simple terms, an agent's trust in another can be understood as a belief that the latter's behavior will support the agent's plans. Subtle relationships underlie trust in social and organizational settings [Castelfranchi and Falcone, 1998]. Without detracting from such principles, this paper takes a narrower view of trust: here an agent seeks to establish a belief or disbelief that another agent's behavior is good (thus abstracting out details of the agent's own plans and the social and organizational relationships between the two agents). The model proposed here can, however, be used to capture as many dimensions of trust as needed, e.g., timeliness, quality of service, and so on.

*This research was partially supported by the National Science Foundation under grant ITR-0081742.

For rational agents, trust in a party should be based substantially on evidence consisting of positive and negative experiences with it. This evidence can be collected by an agent locally or via a reputation agency or by following a referral protocol. In such cases, the evidence may be implicit in the trust reports obtained that somehow summarize the evidence. This paper develops a principled evidence-based approach for trust that supports two crucial requirements of multiagent systems:

Dynamism. Practical agent systems face the challenge that trust evolves over time, both as additional information is obtained and as the parties being considered alter their behavior.

Composition. It is clear that trust cannot be trivially propagated. For example, A may trust B who trusts C, but A may not trust C. However, we need to combine trust reports that cannot themselves be perfectly trusted, possibly because of their provenance or the way in which they are obtained.

Traditionally, principled approaches to trust have been difficult to come by because of the above requirements. With few exceptions, current approaches for combining trust reports tend to involve ad hoc formulas, which might be simple to implement but are extremely difficult to understand from a conceptual basis. The common idea underlying a solution that satisfies the above requirements is the notion of *discounting*. Dynamism can be accommodated by discounting over time and composition by discounting over the space of sources (i.e., agents). Others have used discounting before, but without adequate mathematical justification. For instance, Yu and Singh [2002] develop such a discounting approach layered on their (principled) Dempster-Shafer account.

The best multiagent application of the present approach is in the work of Wang and Singh [2006a], who develop an algebra for aggregating trust over graphs understood as webs of trust. Wang and Singh concentrate on their algebra and assume a separate, underlying trust model, which is the one developed here. By contrast, the present paper is neutral about the discounting and aggregation mechanisms, and instead develops a principled evidential trust model that would underlie any such agent system where trust reports are gathered from multiple sources.

Following Jøsang [2001], we understand trust based on the

probability of the probability of outcomes, and adopt his idea of a trust space of triples of *belief* (in a good outcome), *disbelief* (or belief in a bad outcome), and *uncertainty*. Trust in this sense is neutral as to the outcome and is reflected in the *certainty* (i.e., one minus the uncertainty). Thus the following three situations are distinguished:

- Trust in a party (i.e., regarding its being good): belief is high, disbelief is low, and uncertainty is low.
- Distrust in a party: belief is low, disbelief is high, and uncertainty is low.
- Lack of trust in a party (pro or con): uncertainty is high.

However, whereas Jøsang defines certainty in an ad hoc manner, we define certainty based on a well-known statistical measure. Despite the increased subtlety of our definition, it preserves a bijection between trust and evidence spaces, enabling combination of trust reports (via mapping them to evidence). Our definition captures the following key intuitions.

Effect of evidence. Certainty *increases* as evidence increases (for a fixed ratio of positive and negative observations).

Effect of conflict. Certainty *decreases* as the extent of conflict increases in the evidence.

Jøsang's approach satisfies the intuition about evidence but fails the intuition about conflict. It falsely predicts that mounting *conflicting* evidence increases certainty—and equally as much as mounting confirmatory evidence. Say Alice deals with Bob four times or obtains (fully trustworthy) reports about Bob from four witnesses: in either case, her evidence would be between 0 and 4 positive experiences. It seems uncontroversial that Alice's certainty is greatest when the evidence is all in favor or all against and least when the evidence is equally split. Section 3.2 shows that Jøsang assigns the same certainty in each case.

Yu and Singh [2002] model positive, negative, or neutral evidence, and apply Dempster-Shafer theory to compute trust. Neutral experiences yield uncertainty, but conflicting positive or negative evidence doesn't increase uncertainty. Further, for conflicting evidence, Dempster-Shafer theory can yield unintuitive results [Sentz and Ferson, 2002]. Say Pete sees two physicians, Dawn and Ed, for a headache. Dawn says Pete has meningitis, a brain tumor, or neither with probabilities 0.79, 0.2, and 0.01, respectively. Ed says Pete has a concussion, a brain tumor, or neither with probabilities 0.79, 0.2, and 0.01, respectively. Dempster-Shafer theory yields that the probability of a brain tumor is 0.725, which is highly counterintuitive, because neither Dawn nor Ed thought that was likely.

This paper contributes (1) a rigorous, probabilistic definition of certainty that satisfies the above key intuitions, (2) the establishment of a bijection between trust reports and evidence, which enables the principled combination of trust reports, and (3) an efficient algorithm for computing this bijection.

2 Modeling Certainty

The proposed approach is based on the fundamental intuition that an agent can model the behavior of another agent in probabilistic terms. Specifically, an agent can represent the probability of a positive experience with, i.e., good behavior by, another agent. This probability must lie in the real interval $[0, 1]$. The agent's trust corresponds to how strongly the agent believes that this probability is a specific value (whether large or small, we don't care). This strength of belief is also captured in probabilistic terms. To this end, we formulate a probability density function of the probability of a positive experience. Following [Jøsang, 1998], we term this a *probability-certainty density function (PCDF)*. In our approach, unlike Jøsang's, certainty is a statistical measure defined on a PCDF.

2.1 Certainty from a PCDF

Because the cumulative probability of a probability lying within $[0, 1]$ must equal 1, all PCDFs must have the mean density of 1 over $[0, 1]$, and 0 elsewhere. Lacking additional knowledge, a PCDF would be a uniform distribution over $[0, 1]$. However, with additional knowledge, the PCDF would deviate from the uniform distribution. For example, knowing that the probability of good behavior is at least 0.5, we would obtain a distribution that is 0 over $[0, 0.5)$ and 2 over $[0.5, 1]$. Similarly, knowing that the probability of good behavior lies in $[0.5, 0.6]$, we would obtain a distribution that is 0 over $[0, 0.5)$ and $(0.6, 1]$, and 10 over $[0.5, 0.6]$.

In formal terms, let $p \in [0, 1]$ represent the probability of a positive outcome. Let the distribution of p be given as a function $f : [0, 1] \mapsto [0, \infty)$ such that $\int_0^1 f(p)dp = 1$. The probability that the probability of a positive outcome lies in $[p_1, p_2]$ can be calculated by $\int_{p_1}^{p_2} f(p)dp$. The mean value of f is $\frac{\int_0^1 f(p)dp}{1-0} = 1$. When we know nothing else, f is a uniform distribution over probabilities p. That is, $f(p) = 1$ for $p \in [0, 1]$ and 0 elsewhere. This reflects the Bayesian intuition of assuming an equiprobable prior. The uniform distribution has a certainty of 0. As more knowledge is acquired, the probability mass shifts so that $f(p)$ is above 1 for some values of p and below 1 for other values of p.

Our key intuition is that the agent's trust corresponds to increasing deviation from the uniform distribution. Two of the most established measures for deviation are standard deviation and mean absolute deviation (MAD). MAD is more robust, because it does not involve squaring (which can increase standard deviation because of outliers or "heavy tail" distributions such as the notorious Cauchy distribution). Absolute values can sometimes complicate the mathematics. But, in the present setting, MAD turns out to yield straightforward mathematics. In a discrete setting involving data points $x_1 \ldots x_n$ with mean $\hat{x}$, MAD is given by $\frac{1}{n}\Sigma_{i=1}^{n}|x_i - \hat{x}|$. In the present case, instead of n we divide by the size of the domain, i.e., 1. Because a PCDF has a mean value of 1, increase in some parts must match reduction elsewhere. Both increase and reduction from 1 are counted by $|f(p) - 1|$. Definition 1 scales the MAD for f by $\frac{1}{2}$ to remove this double counting.

Definition 1 *The certainty based on f, c_f, is given by* $c_f = \frac{1}{2}\int_0^1 |f(p) - 1|dp$

Certainty captures the fraction of the knowledge that we do have. For motivation, consider randomly picking a ball from a bin that contains N balls colored white or black. Suppose

p is the probability that the ball randomly picked is white. If we have no knowledge about how many white balls there are in the bin, we can't estimate p with any confidence That is, certainty $c = 0$. If we know that exactly m balls are white, then we have perfect knowledge about the distribution. We can estimate $p = \frac{m}{N}$ with $c = 1$. However, if all we know is that at least m balls are white and at least n balls are black (thus $m + n \leq N$), then we have partial knowledge. Here $c = \frac{m+n}{N}$. The probability of drawing a white ball ranges from $\frac{m}{N}$ to $1 - \frac{n}{N}$. We have

$$f(p) = \begin{cases} 0, & [0\frac{m}{N}) \\ \frac{N}{N-m-n} & p \in [\frac{m}{N}, 1 - \frac{n}{N}] \\ 0 & (1 - \frac{n}{N}, 1]. \end{cases}$$

Using Definition 1, we can confirm that $c_f = \frac{m+n}{N}$:

$$\begin{aligned} c_f &= \tfrac{1}{2}\int_0^1 |f(p) - 1|dp \\ &= \tfrac{1}{2}(\int_0^{\frac{m}{N}} 1\ dp + \int_{\frac{m}{N}}^{1-\frac{n}{N}} (\tfrac{N}{N-m-n} - 1)dp + \int_{1-\frac{n}{N}}^1 1\ dp \\ &= \tfrac{1}{2}(\tfrac{m}{N} + \tfrac{N-m-n}{N}(\tfrac{N}{N-m-n} - 1) + \tfrac{n}{N}) \\ &= \tfrac{m+n}{N} \end{aligned}$$

2.2 Evidence and Trust Spaces Conceptually

For simplicity, we model a (rating) agent's experience with a (rated) agent as a binary event: positive or negative. Evidence is conceptualized in terms of the numbers of positive and negative experiences. In terms of direct observations, these numbers would obviously be whole numbers. However, our motivation is to combine evidence in the context of trust. As Section 1 motivates, for reasons of dynamism or composition, the evidence may need to be discounted to reflect the aging of or the imperfect trust placed in the evidence source. Intuitively, because of such discounting, the evidence is best understood as if there were real (i.e., not necessarily natural) numbers of experiences. Accordingly, we model the evidence space as $E = \mathbb{R} \times \mathbb{R}$, a two-dimensional space of reals. The members of E are pairs $\langle r, s\rangle$ corresponding to the numbers of positive and negative experiences, respectively. Combining evidence is trivial: simply perform vector sum.

Definition 2 *Define evidence space* $E = \{(r, s)|r \geq 0, s \geq 0, t = r + s > 0\}$

Let x be the probability of a positive outcome. The posterior probability of evidence $\langle r, s\rangle$ is the conditional probability of x given $\langle r, s\rangle$ [Casella and Berger, 1990, p. 298].

Definition 3 *The conditional probability of* x *given* $\langle r, s\rangle$ *is*

$$\begin{aligned} f(x|\langle r, s\rangle) &= \frac{g(\langle r,s\rangle|x)f(x)}{\int_0^1 g(\langle r,s\rangle|x)f(x)dx} \\ &= \frac{x^r(1-x)^s}{\int_0^1 x^r(1-x)^s dx} \end{aligned}$$

where $g(\langle r, s\rangle|x) = \begin{pmatrix} r+s \\ r \end{pmatrix} x^r(1-x)^s$

Traditional probability theory models the event $\langle r, s\rangle$ by $(\alpha, 1-\alpha)$, the expected probabilities of positive and negative outcomes, respectively, where $\alpha = \frac{r+1}{r+s+2}$. The traditional probability model ignores uncertainty.

A trust space consists of *trust reports* modeled in a three-dimensional space of reals in $(0, 1)$. Each point in this space is a triple $\langle b, d, u\rangle$, where $b + d + u = 1$, representing the weights assigned to belief, disbelief, and uncertainty, respectively. Certainty c is simply $1 - u$. Thus $c = 1$ and $c = 0$ indicate perfect knowledge and ignorance, respectively.

Combining trust reports is nontrivial. Our proposed definition of certainty is key in accomplishing a bijection between evidence and trust reports. The problem of combining independent trust reports is reduced to the problem of combining the evidence underlying them. (Definitions 2 and 4 are based on [Jøsang, 2001].)

Definition 4 *Define trust space as* $T = \{(b, d, u)|b > 0, d > 0, u > 0, b + d + u = 1\}$.

2.3 From Evidence to Trust Reports

Using Definition 3, define certainty based on evidence $\langle r, s\rangle$:

Definition 5 $c(r, s) = \frac{1}{2}\int_0^1 |\frac{(x^r(1-x)^s}{\int_0^1 x^r(1-x)^s dx} - 1|dx$

Throughout, r, s, and $t = r + s$ refer to positive, negative, and total evidence, respectively. Importantly, $\alpha = \frac{r+1}{t+2}$, the expected value of the probability of a positive outcome, also characterizes *conflict* in the evidence. Clearly, $\alpha \in (0, 1)$: α approaching 0 or 1 indicates unanimity, whereas $\alpha = 0.5$ means $r = s$, i.e., maximal conflict in the body of evidence. We can write $c(r, s)$ as $c((t + 2)\alpha - 1, (t + 2)(1 - \alpha) - 1)$. When α is fixed, certainty is a function of t, written $c(t)$. When t is fixed, it is a function of α, written $c(\alpha)$. And, $c'(t)$ and $c'(\alpha)$ are the corresponding derivatives.

The following is our transformation from evidence to trust spaces. This transformation relates positive and negative evidence to belief and disbelief, respectively, but discounted by the certainty. The idea of adding 1 each to r and s (and thus 2 to $r + s$) follows Laplace's famous *rule of succession* for applying probability to inductive reasoning [Sunrise, 2006]. This reduces the impact of sparse evidence, and is sometimes termed *Laplace smoothing*. If you only made one observation and it was positive, you would not want to conclude that there would never be a negative observation. As the body of evidence increases, the increment of 1 becomes negligible. More sophisticated formulations of rules of succession exist [Ristad, 1995], but Laplace's rule is simple and reasonably effective for our present purposes. Laplace's rule is insensitive to the number of outcomes in that 1 is always added. The effect of this statistical "correction" (the added 1) decreases inversely as the number of outcomes being considered increases. More sophisticated approaches may be thought of as decreasing the effects of their corrections more rapidly.

Definition 6 *Let* $Z(r, s) = (b, d, u)$ *be a transformation from* E *to* T *such that* $Z = (b(r, s), d(r, s), u(r, s))$, *where* $b(r, s) = \alpha c(r, s)$, $d(r, s) = (1 - \alpha)c(r, s)$, *and* $u(r, s) = 1 - c(r, s)$.

One can easily verify that $c(0, 1) > 0$. In general, because $t = r + s > 0$, $c(r, s) > 0$. Moreover, $c(r, s) < 1$: thus, $1 - c(r, s) > 0$. This coupled with the rule of succession ensures that $b > 0$, $d > 0$, and $u > 0$. Notice that $\alpha = \frac{b}{b+d}$.

Jøsang *et al.* [1998] map evidence $\langle r, s\rangle$ to a trust triple $(\frac{r}{t+1}, \frac{s}{t+1}, \frac{1}{t+1})$. Two main differences with our approach are: (1) they ignore the rule of succession and (2) in essence, they define certainty as $\frac{t}{t+1}$. They offer no mathematical justification for doing so. Section 3.2 shows an unintuitive consequence of their definition.

3 Important Properties and Computation

We now show that the above definition yields important formal properties and how to compute with it.

3.1 Increasing Experiences with Fixed Conflict

Consider the scenario where the total number of experiences increases for fixed $\alpha = 0.70$. For example, compare observing 6 good episodes out of 8 with observing 69 good episodes out of 98. The expected value, α, is the same in both cases, but the certainty is clearly greater in the second. In general, we would expect certainty to increase as the amount of evidence increases. Definition 5 yields a certainty of 0.46 from $\langle r, s\rangle = \langle 6, 2\rangle$, but a certainty of 0.70 for $\langle r, s\rangle = \langle 69, 29\rangle$.

Figure 1 plots how certainty varies with t when $\alpha = 0.5$. Theorem 1 captures this property in general.

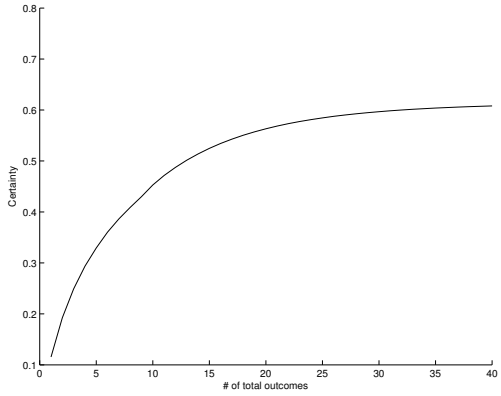

Figure 1: Certainty increases with t when conflict ($\alpha = 0.5$) is fixed; X-axis: t; Y-axis: $c(t)$

Theorem 1 *Fix α. Then $c(t)$ increases with t for $t > 0$.*

Proof idea: Show that $c'(t) > 0$ for $t > 0$.

The full proofs of this and other theorems of this paper are included in a technical report [Wang and Singh, 2006b].

3.2 Increasing Conflict with Fixed Experience

Another important scenario is when the total number of experiences is fixed, but the evidence varies to reflect different levels of conflict by using different values of α. Clearly, certainty should increase as r or s dominates the other (i.e., α approaches 0 or 1) but should reduce as r and s are balanced (i.e., α approaches 0.5). Figure 2 plots certainty for fixed t and varying conflict.

More specifically, consider Alice's example from Section 1. Table 1 shows the effect of conflict where $t = 4$.

Theorem 2 captures the property that certainty increases with increasing unanimity.

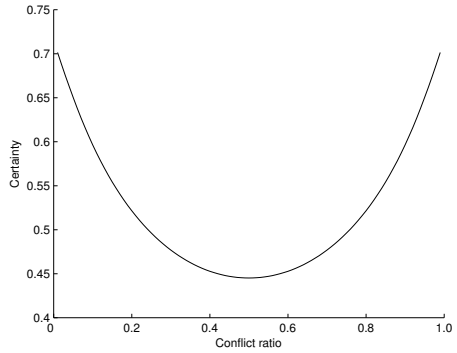

Figure 2: Certainty is concave when t is fixed at 10; X-axis: $r + 1$; Y-axis: $c(\alpha)$; minimum occurs at $r = s = 5$

Table 1: Certainty computed by different approaches for varying conflict

	$\langle 0, 4\rangle$	$\langle 1, 3\rangle$	$\langle 2, 2\rangle$	$\langle 3, 1\rangle$	$\langle 4, 0\rangle$
Our approach	0.54	0.35	0.29	0.35	0.54
Jøsang et al.	0.80	0.80	0.80	0.80	0.80
Yu & Singh	0	0	0	0	0

Theorem 2 *$c(\alpha)$ is decreasing when $0 < \alpha \leq \frac{1}{2}$, increasing when $\frac{1}{2} \leq \alpha < 1$ and $c(\alpha)$, and minimum at $\alpha = \frac{1}{2}$.*

Proof idea: Show that $c'(\alpha) < 0$ when $\alpha \in [0, 0.5)$ and $c'(\alpha) > 0$ when $\alpha \in [0.5, 1.0)$.

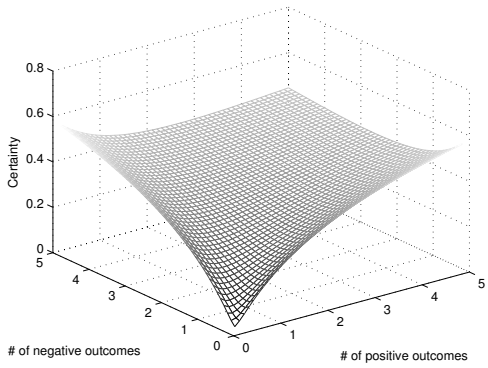

Figure 3: X-axis: r; Y-axis: s; Z-axis: certainty

Putting the above results together suggests that the relationship between certainty on the one hand and positive and negative evidence on the other is nontrivial. Figure 3 confirms this intuition by plotting certainty against r and s as a surface.

3.3 Bijection Between Evidence and Trust Reports

The ability to combine trust reports effectively relies on being able to map between the evidence and the trust spaces. With such a mapping in hand, to combine two trust reports, we would simply perform the following steps: (1) map trust reports to evidence; (2) combine the evidence; (3) transform

the combined evidence to a trust report. The following theorem establishes that Z has a unique inverse Z^{-1}.

Theorem 3 *The transformation Z is a bijection.*

Proof sketch: Given $(b, d, u) \in T$, we need $(r, s) \in E$ such that $Z(r, s) = (b, d, u)$. As explained in Section 2.3, $\alpha = \frac{b}{b+d}$. Thus, we only need to find t such that $c(t) = 1 - u$. The existence and uniqueness of t is proved by showing that

1. $c(t)$ is increasing when $t > 0$ (Theorem 1)
2. $\lim_{t\to\infty} c(t) = 1$
3. $\lim_{t\to 0} c(t) = 0$

Briefly, Yu and Singh [2002] base uncertainty not on conflict, but on intermediate (neither positive not negative) outcomes. Let's revisit Pete's example of Section 1. In our approach, Dawn and Ed's diagnoses correspond to two b, d, u triples (where b means "tumor" and d means "not a tumor"): $(0.2, 0.79, 0.01)$ and $(0.2, 0.79, 0.01)$, respectively. Combining these we obtain the b, d, u triple of $(0.21, 0.78, 0.01)$. That is, the weight assigned to a tumor is 0.21 as opposed to 0.725 by Dempster-Shafer theory, which is unintuitive, because a tumor is Dawn and Ed's least likely prediction.

3.4 Algorithm and Complexity

No closed form is known for Z^{-1}. Algorithm 1 calculates Z^{-1} (via binary search on $c(t)$) to any necessary precision, $\epsilon > 0$. Here $t_{max} > 0$ is the maximum evidence considered.

```
1 α = b/(b+d);
2 t1 = 0;
3 t2 = tmax;
4 while t2 − t1 ≥ ε do
5 |   t = (t1+t2)/2;
6 |   if c(t) < c then t1 = t else t2 = t
7 return r = ((t + 2)α − 1), s = t − r
```

Algorithm 1: Calculating $(r, s) = Z^{-1}(b, d, u)$

Theorem 4 *The complexity of Algorithm 1 is $\Omega(-\lg \epsilon)$.*

Proof: After the **while** loop iterates i times, $t_2 - t_1 = t_{max}2^{-i}$. Eventually, $t_2 - t_1$ falls below ϵ, thus terminating the **while** loop. Assume it terminates in n iterations. Then, $t_2 - t_1 = t_{max}2^{-n} < \epsilon \leq t_{max}2^{-n+1}$. This implies $2^n > \frac{t_{max}}{\epsilon} \geq 2^{n-1}$. That is, $n > (\lg t_{max} - \lg \epsilon) \geq n - 1$.

4 Discussion

This paper is meant to offer a theoretical development of trust that would underlie a variety of situations where trust reports based on evidence are combined. In particular, it contributes to a mathematical understanding of trust, especially as it underlies a variety of multiagent applications. These include referral systems and webs of trust in particular, in studying which we identified the need for this research. Such applications require a natural treatment of composition and discounting in an evidence-based framework.

Further, an evidence-based notion of trust must support important properties regarding the effects of increasing evidence (for constant conflict) and of increasing conflict (for constant evidence). The theoretical validation provided here is highly valuable in a general-purpose conceptually driven mathematical approach. The main technical insight of this paper is how to manage the duality between trust and evidence spaces in a manner that provides a rigorous basis for combining trust reports.

Let's briefly revisit the topic of trust dynamics from Section 1. The foregoing showed how trust evolves with respect to increasing outcomes under different conditions. The same properties apply to the evolution of trust over time, that is, as time passes and more evidence is obtained. A crucial observation is that because of the bijection we established, the historical evidence at any point can be summarized in a belief-disbelief-uncertainty triple. New evidence can then be added as explained above. Moreover, we can discount the value of evidence over time if necessary, e.g., at every time step (chosen based on the domain: every hour or day, or after every transaction). Thus new evidence would have a greater impact than older evidence.

A payoff of this approach is that an agent who wishes to achieve a specific level of certainty can compute how much evidence would be needed at different levels of conflict. Or, the agent can iteratively compute certainty to see if it has reached an acceptable level.

4.1 Directions

This work has opened up some important directions for future work. An important technical challenge is to extend the above work from binary to multivalued events. Such an extension will enable us to handle a larger variety of interactions among people and services. A current direction is to experimentally validate this work, doing which is made difficult by the lack of established datasets and testbeds, but the situation is improving in this regard [Fullam *et al.*, 2005].

4.2 Literature on Trust

A huge amount of research has been conducted on trust, even if we limit our attention to evidential approaches. Abdul-Rahman and Hailes [2000] present an early model for computing trust. However, their approach is highly ad hoc and limited. Specifically, various weights are simply added up without any mathematical justification. Likewise, the term *uncertainty* is described but without any foundation.

The Regret system combines several aspects of trust, notably the social aspects [Sabater and Sierra, 2002]. It involves a number of formulas, which are given intuitive, but not mathematical, justification. A lot of other work, e.g., [Huynh *et al.*, 2006], involves heuristics that combine multiple information sources to judge trust. It would be an interesting direction to combine a rigorous approach such as ours with the above heuristic approaches to capture a rich variety of practical criteria well.

Teacy *et al.* [2005] develop a probabilistic treatment of trust. They model trust in terms of confidence that the expected value lies within a specified error tolerance. An agent's confidence increases with the error tolerance. Teacy *et*

al. study combinations of probability distributions to which the evaluations given by different agents might correspond. They do not formally study certainty. And their approach doesn't yield a probabilistically valid method for combining trust reports.

4.3 Literature on Information Theory

Shannon entropy [1948] is the best known information-theoretic measure of uncertainty. It is based on a discrete probability distribution $p = \langle p(x)|x \in X\rangle$ over a finite set X of alternatives (elementary events). Shannon's formula encodes the number of bits required to obtain certainty: $S(p) = -\sum_{x\in X} p(x)\log_2 p(x)$. Here $S(p)$ can be viewed as the weighted average of the conflict among the evidential claims expressed by p. More complex, but less well-established, definitions of entropy have been proposed for continuous distributions as well, e.g., [Smith, 2001].

Entropy, however, is not suitable for the present purposes of modeling evidential trust. Entropy models bits of missing information which ranges from 0 to ∞. At one level, this disagrees with our intuition that, for the purposes of trust, we need to model the confidence placed in a probability estimation. Moreover, the above definitions cannot be used in measuring the uncertainty of the probability estimation based on past positive and negative experiences.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and Chung-Wei Hang for useful discussions.

References

[Abdul-Rahman and Hailes, 2000] Alfarez Abdul-Rahman and Stephen Hailes. Supporting trust in virtual communities. In *Proceedings of the 33rd Hawaii International Conference on Systems Science*, 2000.

[Casella and Berger, 1990] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Press, 1990.

[Castelfranchi and Falcone, 1998] Cristiano Castelfranchi and Rino Falcone. Principles of trust for MAS: cognitive anatomy, social importance, and quantification. In *Proceedings of the 3rd International Conference on Multiagent Systems*, pages 72–79, 1998.

[Fullam *et al.*, 2005] Karen Fullam, Tomas B. Klos, Guillaume Muller, Jordi Sabater, Andreas Schlosser, Zvi Topol, K. Suzanne Barber, Jeffrey S. Rosenschein, Laurent Vercouter, and Marco Voss. A specification of the Agent Reputation and Trust (ART) testbed: experimentation and competition for trust in agent societies. In *Proceedings of the 4th International Joint Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 512–518. ACM Press, July 2005.

[Huynh *et al.*, 2006] Trung Dong Huynh, Nicholas R. Jennings, and Nigel R. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Journal of Autonomous Agents and MultiAgent Systems*, 13(2):119–154, September 2006.

[Jøsang, 1998] Audun Jøsang. A subjective metric of authentication. In *Proceedings of the 5th European Symposium on Research in Computer Security (ESORICS)*, volume 1485 of *LNCS*, pages 329–344. Springer-Verlag, 1998.

[Jøsang, 2001] Audun Jøsang. A logic for uncertain probabilities. *Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9:279–311, 2001.

[Ristad, 1995] Eric Sven Ristad. A natural law of succession. TR 495-95, Department of Computer Science, Princeton University, July 1995.

[Sabater and Sierra, 2002] Jordi Sabater and Carles Sierra. Reputation and social network analysis in multi-agent systems. In *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 475–482. ACM Press, July 2002.

[Sentz and Ferson, 2002] Karl Sentz and Scott Ferson. Combination of evidence in Dempster Shafer theory. TR 0835, Sandia National Laboratories, Albuquerque, New Mexico, 2002.

[Shannon, 1948] Claude E. Shannon. The mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[Smith, 2001] Jonathan D. H. Smith. Some observations on the concepts of information-theoretic entropy and randomness. *Entropy*, 3:1–11, 2001.

[Sunrise, 2006] The sunrise problem, 2006. http://en.wikipedia.org/wiki/Sunrise_problem.

[Teacy *et al.*, 2005] Luke Teacy, Jigar Patel, Nicholas Jennings, and Michael Luck. Coping with inaccurate reputation sources: Experimental analysis of a probabilistic trust model. In *Proceedings of the 4th International Joint Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 997–1004. ACM Press, July 2005.

[Wang and Singh, 2006a] Yonghong Wang and Munindar P. Singh. Trust representation and aggregation in a distributed agent system. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, pages 1425–1430, 2006a.

[Wang and Singh, 2006b] Yonghong Wang and Munindar P. Singh. Trust via evidence combination: A mathematical approach based on certainty. TR 2006-11, North Carolina State University, Raleigh, 2006b.

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