TRIBE: Trust Revision for Information Based on Evidence

Murat Şensoy*†, Geeth de Mel‡§, Lance Kaplan‡, Tien Pham‡, and Timothy J. Norman†
* Department of Computer Science, Ozyegin University, Istanbul, Turkey.
† Department of Computing Science, University of Aberdeen, UK.
‡ US Army Research Lab, Adelphi, MD 20783, USA.
§ IBM T. J. Watson Research Center, Hawthorne, NY, USA.

***Abstract*—In recent years, the number of information sources available to support decision-making has increased dramatically. However, more information sources do not always mean higher precision in the fused information. This is partially due to the fact that some of these sources may be erroneous or malicious. Therefore, it is critical to asses the trust in information before performing fusion. To estimate trust in information, existing approaches use trustworthiness of its source as a proxy. We argue that conflicts between information may also serve as evidence to reduce trust in information. In this paper, we use subjective opinions to represent information from diverse sources. We propose to exploit conflicts between opinions to revise their trustworthiness. For this purpose, we formalise trust revision as a constraint optimisation problem. Through extensive empirical studies, we show that our approach significantly outperform existing ones in the face of malicious information sources.**

Keywords*—*Information Fusion, Trust, Constraint Optimisation, Subjective Logic, Dempster-Shafer Theory of Evidence.

I. Introduction

The volumes of information streamed and collected from disparate sensors and information sources have increased dramatically in recent years. Fusing such multimodal information may increase decision makers' ability to make informed decision in rapidly changing complex environments [8]. However, information from some of these sources may be noisy, incomplete, and even misleading. If the misleading information is not discounted beforehand, it may takeover the useful information, thus making the fused information misleading as well. Therefore, it is very important to estimate trustworthiness of information before fusion and discount or eliminate the information which is likely to mislead.

A common and intuitive way of estimating trust in information is to use trustworthiness of its source as a proxy. The trustworthiness of sources is an important input while estimating trust in information they provide. However, trustworthiness of information may not always correlate well with the trustworthiness of its source. A prime example for this is the case where information from two or more equally trustworthy sources is in conflict. These conflicts can be resolved by revising trust in the information. That is, the information from a trustworthy source may be regarded as untrustworthy if it conflicts with information from other sources.

When we consider information from a single source, we usually have only the trustworthiness of its source as evidence for its trustworthiness. However, consideration of multiple sources reveals that information can be conflicting. These conflicts between information provided by the sources can be viewed as an additional evidence beyond source trustworthiness while estimating trust in the information.

In this paper, we use subjective opinions to represent information. Subjective opinions express subjective beliefs about the truth of propositions with degrees of uncertainty. They are based on the belief functions of Dempster-Shafer theory of evidence [12] and formalised within Subjective Logic [5]. We use a well-known statistical model to estimate trust in information sources based on the available evidence. Then, we propose a definition of conflict between subjective opinions about the same proposition. The conflicts between opinions are used as evidence to necessitate the trust revisions on these opinions. We formalise trust revision as a constraint optimisation problem and show through extensive simulations that our approach significantly improves the precision of information fusion and outperforms existing approaches.

The remainder of the paper is structured as follows. We begin by introducing subjective opinions in Section II. Section III describes – TRIBE – our approach for trust revision for information based on evidence, where we describe how trust in information sources are estimated, how conflicts between opinions are detected, and how the detected conflicts can be exploited to revise trust in opinions before performing fusion. We evaluate our approach in Section IV, and discuss it with respect to the existing work and draw future research directions in Section V. Lastly, we conclude our paper in Section VI.

II. Subjective Opinions

Dempster-Shafer Theory (DST) offers means to characterise an agent's view of the state of world by assigning *basic probability masses* to subsets of truth assignments of propositions in the logic. In this paper, we adopt Subjective Logic (SL) proposed by Jøsang [5], which can be considered as an interpretation and extension of DST. In SL, a *binomial opinion* about a binary proposition x is represented by a triple

Research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorised to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon. Dr. Şensoy thanks to The Scientific and Technological Research Council of Turkey (TUBITAK) for its support under grant 111K476.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUL 2013		**00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
TRIBE: Trust Revision for Information Based on Evidence	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **US Army Research Lab,Adelphi,MD,20783**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the 16th International Conference on Information Fusion held in Istanbul, Turkey on 9-12 July 2013. Sponsored in part by Office of Naval Research Global.

14. ABSTRACT
In recent years, the number of information sources available to support decision-making has increased dramatically. However, more information sources do not always mean higher precision in the fused information. This is partially due to the fact that some of these sources may be erroneous or malicious. Therefore, it is critical to asses the trust in information before performing fusion. To estimate trust in information, existing approaches use trustworthiness of its source as a proxy. We argue that conflicts between information may also serve as evidence to reduce trust in information. In this paper, we use subjective opinions to represent information from diverse sources. We propose to exploit conflicts between opinions to revise their trustworthiness. For this purpose, we formalise trust revision as a constraint optimisation problem. Through extensive empirical studies, we show that our approach significantly outperform existing ones in the face of malicious information sources.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**8**	

$w_x = (b_x, d_x, u_x)$ which is derived from the basic probability masses assigned to subsets of truth assignments. In the opinion w_x, b_x, also denoted by $b(w_x)$, is the belief about x — the summation of the probability masses that entail x; d_x, also denoted by $d(w_x)$, is the disbelief about x — the summation of the probability masses that entail $\neg x$; and u_x, also denoted by $u(w_x)$, is the uncertainty about x — the summation of the probability masses that neither entail x nor entail $\neg x$. The constraints over the probability mass assignment function require that $b_x + d_x + u_x = 1$ and $b_x, d_x, u_x \in [0, 1]$. When a more concise notation is required, we use (b_x, d_x) instead of (b_x, d_x, u_x), since $u_x = 1 - b_x - d_x$.

Using subjective opinions, true/false values in binary logic or membership values in fuzzy logic can be represented. Let $dangerous(zone2)$ be a proposition. The opinion $(1, 0, 0)$ indicates that the proposition is *true*; while $(0, 1, 0)$ implies that it is *false*. Similarly, an opinion like $(0.6, 0.4, 0.0)$ implies that the proposition is both *true* and *false* with different belief masses. Furthermore, the uncertainty in an opinion may imply incomplete knowledge or lack of evidence about the proposition. The negation over an opinion w_x is defined as $\neg(b_x, d_x, u_x) = (d_x, b_x, u_x) = (b_{\neg x}, d_{\neg x}, u_{\neg x})$. For example, $\neg(1, 0, 0) = (0, 1, 0)$.

An agent i's opinion about a proposition x is denoted by $w_x^i = (b_x^i, d_x^i, u_x^i)$. This opinion may not be *directly* used by another agent j. Agent j could have a view of the reliability or competence of i with respect to x. Shafer [9] and Jøsang [5] proposed a discounting operator $\otimes$ to normalise the belief and disbelief in w_x^j based on t_i^j, i.e., the degree of trust j has of i. The discounted opinion, $w_x^{j:i}$, is computed as:

$$w_x^{j:i} = w_x^i \otimes t_i^j = (b_x^i \times t_i^j, d_x^i \times t_i^j)$$

In this paper, we omit the agent superscript if it does not lead to any confusion.

Following Jøsang [5], we assume that opinions are formed on the basis of positive and negative evidence. Let r and s be the number of positive and negative evidences (i.e., equivalently weighted pieces of evidence) about the proposition x, respectively. Then, an opinion composed of b, d, and u is computed based on evidence r and s as in Equation 1.

$$op(r, s) = (b, d, u) = \left(\frac{r}{r+s+2}, \frac{s}{r+s+2}, \frac{2}{r+s+2}\right) \quad (1)$$

There is a bijection between the opinion (b, d, u) and associated evidence (r, s). Given an opinion, amounts of positive and negative evidence are computed as in Equation 2.

$$r = \frac{2 \times b}{u} \text{ and } s = \frac{2 \times d}{u} \quad (2)$$

In this paper, we use $r(w)$ and $s(w)$ as the functions that map opinion w to positive and negative evidence (i.e., r and s), respectively, based on Equation 2. Each opinion for a binary proposition is associated with a base rate a, which represents the a priority probability that the proposition is *true*. Using the base rate a for the opinion w, the opinion's probability expectation value can be computed using Equation 3 [5].

$$E(w, a) = b(w) + a \times u(w) = \frac{r(w) + a \times 2}{r(w) + s(w) + 2} \quad (3)$$

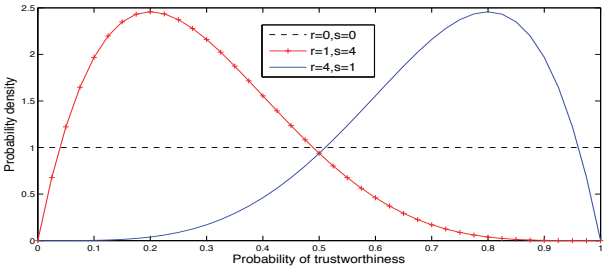

Fig. 1. Beta PDFs for different amounts of evidence.

III. Trust-based Fusion of Opinions

In sensing applications, an information consumer may receive many different opinions from diverse sources with different reliabilities. The consumer should discount these opinions based on the trustworthiness of the respective sources and then fuse these discounted opinions using a fusion operator. Using the fusion and discounting operators, an agent j can fuse information based on the trustworthiness of the sources as shown below, where t_i^j is j's trust in i_i in the context.

$$\begin{aligned} &fuse_j(w_y^{i_0}, ..., w_y^{i_n}) = \\ &fuse((w_y^{i_0} \otimes t_{i_0}^j), ..., (w_y^{i_n} \otimes t_{i_n}^j)) \end{aligned} \quad (4)$$

There are a number of fusion operators that an information consumer may use. These operators are out of the scope of this paper, but a more detailed discussion of these operators can be found elsewhere [5], [12].

A. Estimating Trust in Information Sources

The trustworthiness of information sources can be modelled using beta probability density functions [6], [10], [15]. A beta distribution has two parameters $(r_i^j + 1, s_i^j + 1)$, where r_i^j is the amount of positive evidence and s_i^j is the amount of negative evidence for the trustworthiness an agent j has for another agent i. The degree of trust – i.e., t_i^j – is then computed as the expectation value of the beta distribution:

$$t_i^j = \frac{r_i^j + 1}{r_i^j + s_i^j + 2} \quad (5)$$

Figure 1 shows how the beta distributions are shaped in the face of varying amounts of evidence.

In Equation 6, the trust model is extended with the *base rate* for the source i (i.e., a_i^j). Here, the base rate (i.e., *a priori trust value*) serves as a *bias* for or against the trustworthiness of the source [5].

$$t_i^j = \frac{r_i^j + a_i^j \times 2}{r_i^j + s_i^j + 2} \quad (6)$$

The base rate a_i^j allows the agent j to estimate trust in i when there is no evidence. This model is reduced to the trust model in Equation 5 if a_i^j is set to 0.5.

We assume that information consumers have some mechanisms to observe or obtain evidence about the trustworthiness of information sources. These mechanisms are out of the scope of this paper, but examples of such mechanisms can be found else where [6], [10], [15].

B. Conflicts between Opinions

Subjective opinions about the same proposition are generated independently by different sources. When an information consumer gets two or more opinions, these opinions may conflict. For example, Pete sees two physicians Dawn and Ed, for a headache. Dawn says Pete has a *brain tumor* with opinion $(1, 0, 0)$, which means Pete absolutely has a brain tumor. On the other hand, Ed says Pete has a *brain tumor* with opinion $(0, 1, 0)$, which means Pete absolutely does not have a brain tumor. It is easy to conclude that these opinions are in conflict. That is, these strictly certain opinions correspond to logical *true* and *false*, respectively; hence they both cannot be valid at the same time.

The conflict indicates that *at least* one of these opinions is misleading. To resolve this conflict, these opinions may be discounted (possibly at different rates), hence the uncertainty within them is increased enough to resolve the conflict. For instance, discounting the first opinion $(1, 0, 0)$ with 0.0 makes it $(0, 0, 1)$, which absolutely does not conflict with $(0, 1, 0)$ or any other opinion, since it does not contain any belief or disbelief, but only uncertainty. Similarly, discounting both of these opinions with 0.5 leads to opinions $(0.5, 0, 0.5)$ and $(0, 0.5, 0.5)$, which are quite uncertain and may not conflict. Let us note that discounting an opinion with a constant c $(0 \leq c \leq 1)$ implies that the trust in the opinion is c.

In this paper, we argue that conflicts between opinions may serve as evidence for the necessity to increase uncertainty within them. Although there can be more than one way to define conflicts between subjective opinions, in this paper, we introduce Definition 1 where conflicts are defined based on the satisfiability of *beliefs* and *disbeliefs* within opinions.

Definition 1: Let $O = \{w^0, w^1, \ldots, w^n\}$ be opinions about the same proposition, where each opinion $w^i = (b^i, d^i, u^i)$. These opinions are consistent if it is possible to have a valid opinion that can satisfy all of these opinions. An opinion $w^* = (b^*, d^*, u^*)$ can satisfy the opinion $w^i \in O$ *iff* $b^i \leq b^*$ and $d^i \leq d^*$. Although there can be infinitely many such w^*, the one with the highest uncertainty (i.e., one with the highest u^*) would be

$$(max(b^0, b^1, \ldots, b^n),\ max(d^0, d^1, \ldots, d^n)).$$

Therefore, there cannot be any opinion that can satisfy all opinions in O if and only if

$$max(b^0, b^1, \ldots, b^n) + max(d^0, d^1, \ldots, d^n) > 1.$$

That is, O is inconsistent *iff* $\exists w^i, w^j \in O$ s.t. $b^i + d^j > 1$. ∎

The intuition behind the definition of conflicts between opinions is as follows. Let the ground truth about a proposition x be represented as an opinion $w_x^\tau = (b_x^\tau, d_x^\tau)$. Also, let $w_x^i = (b_x^i, d_x^i, u_x^i)$ be an arbitrary opinion about x. If w_x^i has a higher *belief* than w_x^τ does (i.e., $b_x^\tau < b_x^i$), then w_x^i is misleading. Similarly, if w_x^i has a higher *disbelief* than w_x^τ does (i.e., $d_x^\tau < d_x^i$), then w_x^i is misleading. How much w_x^i is misleading depends on how much extra belief and disbelief it imposes. On the other hand, if $b_x^i < b_x^\tau$ and $d_x^i < d_x^\tau$, then w_x^i is not misleading, because it does not impose any extra belief or disbelief, but only extra uncertainty that conflicts with neither the *belief* nor the *disbelief* within the ground truth w_x^τ. In reality, we do not have the ground truth about x. Therefore, we cannot say whether w_x^i is misleading or not. However, if we have another opinion $w_x^j = (b_x^j, d_x^j, u_x^j)$, we can reason about if it is possible to have a ground truth for which neither w_x^i nor w_x^j is misleading. Such a ground truth exists if and only if $max(b_x^i, b_x^j) + max(d_x^i, d_x^j) \leq 1$. If such a ground truth cannot exist, we say w_x^i and w_x^i are in conflict, i.e., at least one of them must be misleading at some degree.

Let us describe our notion of conflicts through a simple example. Jane wants to buy a car from a dealer, but she has limited knowledge about car dealers. Bob is a car dealer, which provides good service with probability 0.8 and bad service with probability 0.2. Therefore, the ground truth about Bob's trustworthiness as a dealer is $(0.8, 0.2, 0)$. Jane asks opinions of David and Jack about Bob's trustworthiness as a dealer. David gives $(0.7, 0.1, 0.2)$ and Jack gives $(0.2, 0.4, 0.4)$. Let us assume that Jane's trust in both David and Jack is 1.0, so she does not discount their opinions before fusion. Based on Definition 1, these opinions are in conflict, because $0.7 + 0.4 > 1$; hence at least one of these opinions must be misleading. At this point, Jane does not know which of these opinions is misleading. Once Jane learns or estimates the ground truth based on her own observations, she may regard $(0.2, 0.4, 0.4)$ as a negative evidence for the trustworthiness of Jack. Similarly, $(0.7, 0.1, 0.2)$ may be regarded as a positive evidence for the trustworthiness of David.

Let us note that if David and Jack gave $(0.3, 0.6)$ and $(0.2, 0.4)$ as their opinions, there would be no conflict, since it is possible to have an opinion such as $(0.3, 0.6)$ that can satisfy both of these opinions. However, these opinions might serve as negative evidences for David and Jack in future, once Jane got to know the ground truth or a good estimate of it.

C. Exploiting Conflicts to Revise Trust in Opinions

Assume that the agent j has received opinions $w^i = (b^i, d^i)$ and $w^k = (b^k, d^k)$ about the same proposition from the sources i and k. Let a_i and $\langle r_i, s_i \rangle$ refer to the a priori trust for i, and the amount of positive and negative evidence that j has for the trustworthiness of i, respectively. Before fusing these opinions, j discounts them using t_i and t_k, which are j's trust in the sources. The discounted opinions are

$$w^{j:i} = (b^{j:i}, d^{j:i}) = (b^i \times t_i, d^i \times t_i) \text{ and}$$

$$w^{j:k} = (b^{j:k}, d^{j:k}) = (b^k \times t_k, d^k \times t_k).$$

If $b^{j:i} + d^{j:k} > 1$ or $b^{j:k} + d^{j:i} > 1$, the discounted opinions are in conflict. We denote the portion of $w^{j:i}$ that conflicts with $w^{j:k}$ as c_{ik}, and refer to it as the *conflicting portion*. For instance, if $b^{j:i} + d^{j:k} > 1$, then $c_{ik} = b^{j:i}$ and $c_{ki} = d^{j:k}$. This conflict implies that j could not successfully estimate trust in these opinions. Hence, w^i and w^k should be further discounted by j to resolve the conflict.

Let us consider an example where $w^i = (0, 1)$ and $w^k = (1, 0)$; also $a_i = a_k = 0.5$, $r_i = 4$, $s_i = 0$, $r_k = 403$, and $s_k = 100$; so, $t_i = 0.83$ and $t_k = 0.8$ based on Equation 6. As a result, discounted opinions are $w^{j:i} = (0, 0.83)$ and $w^{j:k} = (0.8, 0)$. These opinions are in conflict, since $0.8 + 0.83 > 1$, with $c_{ik} = 0.83$ and $c_{ki} = 0.8$.

When w^i and w^k are received by the agent j, they are discounted by t_i and t_k, respectively. However, as explained before, additional discounting is required since the discounted opinions $w^{j:i}$ and $w^{j:k}$ are in conflict. Let $0 \leq \alpha_i, \alpha_k \leq 1$ be the discounting factors that will be applied to $w^{j:i}$ and $w^{j:k}$, respectively, to resolve the conflict. Discounting $w^{j:i}$ by α_i implies discounting the original opinion w^i by $t_i \times \alpha_i$. This corresponds to *revising* the trustworthiness of w^i as $t_i \times \alpha_i$ by speculating about the trustworthiness of the source i regarding this single opinion. That is, even though the trustworthiness of i is t_i based on the existing evidence $\langle r_i, s_i \rangle$, it becomes $t_i \times \alpha_i$ for this specific opinion w^i; therefore, $t_i \times \alpha_i$ effectively becomes the trust in w^i from agent j's point of view. Below, we propose a metric to measure how much the agent j needs to speculate about the trustworthiness of i regarding w^i.

To decrease trust from t_i to $t_i \times \alpha_i$, j needs additional negative evidence, which is called *speculative evidence* and designated by ρ_i. Our intuition is that it is less likely for a trustworthy source to present additional negative *speculative evidence* than it is for an untrustworthy one, and thus the receipt of such evidence should be tempered by $(\bar{t}_i)^\kappa$. Here, $\bar{t}_i = 1 - t_i$ represents the *distrust* that agent j has in the source i – i.e., the likelihood that j will receive additional negative evidence given its experiences with the source. The calibration constant $\kappa \geq 0$ enables j to vary the influence that prior experience has on its prediction that a source will present negative evidence in the future. If $\kappa = 0$, for example, j assumes that all sources are equally likely to provide negative evidence. We set κ to 2 in our implementation and examples in this paper. Using the trust model in Equation 6, we obtain:

$$t_i \times \alpha_i = \frac{r_i + a_i \times 2}{r_i + s_i + 2} \times \alpha_i = \frac{r_i + a_i \times 2}{s_i + r_i + 2 + \rho_i.(\bar{t}_i)^\kappa} = \frac{r_i + a_i \times 2}{r_i + s_i + 2 + \rho_i.(\frac{s_i + (1-a_i)\times 2}{r_i + s_i + 2})^\kappa}$$

Rearranging this for ρ_i yields:

$$\rho_i = \frac{\nu_i}{\alpha_i} - \nu_i \quad \text{where} \quad \nu_i = \frac{(r_i + s_i + 2)^{\kappa+1}}{(s_i + (1 - a_i) \times 2)^\kappa} \tag{7}$$

To resolve the conflict, the agent j may use different $\langle \alpha_i, \alpha_k \rangle$ pairs, each of which may lead to different amount of evidence that should be speculated by j. For example, to resolve the conflict in the example above, j can use $\alpha_i = 1$ and $\alpha_k = 17/0.8 \approx 0.21$; so the trust in w^k is revised to 0.17 and hence the opinion from k is discounted to $(0.17, 0)$. The speculated evidence for the resolution is then computed as $\rho_i = 0$ and $\rho_k = 46787$. Alternatively, j can use $\alpha_i = 0.2/0.83 \approx 0.24$ and $\alpha_k = 1$; so the trust in w^i is revised to 0.2 and hence the opinion from i is discounted to $(0, 0.2)$. In this case, the speculated evidence for the resolution is computed as $\rho_i = 680.4$ and $\rho_k = 0$. Given the total speculated evidence for each of these two options, it is more rational for the agent j to choose $\alpha_i = 0.24$ and $\alpha_k = 1$, and revise its trust in w^i as 0.2.

Our approach for resolving conflicts is based on minimizing the total amount of speculative evidence used while revising trust in conflicting opinions. We generalise this approach for any number of opinions with arbitrary conflicts. Let us assume we have a set of conflicting discounted opinions $\{\langle w^{j:i}, w^{j:k} \rangle, \ldots, \langle w^{j:m}, w^{j:n} \rangle\}$ and constants $\{\nu_i, \nu_k, \ldots, \nu_m, \nu_n\}$ that are derived from trust evidence about information sources. To determine the optimum discounting factors $\{\alpha_i, \alpha_k, \ldots, \alpha_m, \alpha_n\}$ for these opinions, we construct the following optimisation problem with a multivariate non-linear objective function and linear constraints.

$$\begin{aligned} &\arg\min_{\overrightarrow{\alpha}} f(\overrightarrow{\alpha}) \quad \text{where} \\ f(\langle \alpha_i, \alpha_k, \ldots, \alpha_m, \alpha_n \rangle) &= \frac{\nu_i}{\alpha_i} + \frac{\nu_k}{\alpha_k} + \ldots + \frac{\nu_m}{\alpha_m} + \frac{\nu_n}{\alpha_n} \\ \text{such that} \quad & 0 \leq \alpha_i \leq 1, \ldots, 0 \leq \alpha_n \leq 1 \\ \text{and} \quad & 0 \leq c_{ik}\alpha_i + c_{ki}\alpha_k \leq 1, \ldots, \\ & 0 \leq c_{mn}\alpha_m + c_{nm}\alpha_n \leq 1. \end{aligned}$$

The objective function presented here is convex. The convex property of the objective function guarantees that any local minima is also the global minimum. Existing non-linear constraint optimisation techniques can be used to solve this problem in order to estimate the best discounting factors.

In this section we have formalised the problem of computing additional discounting factors for *opinions* received about the world from different sources so that we may formulate a consistent set of opinions from which we can draw reliable conclusions. To draw conclusions, these consistent opinions are fused using a fusion operator. This paper does not assume a specific fusion operator for this purpose. However, in the next section, we present the fusion operator used in our implementation.

IV. Evaluation

In the previous sections of the paper, we have described our main contributions. In this section, we provide details of our implementation of the proposed approach. We then evaluate it with respect to similar approaches from the literature through a set of extensive simulations.

A. Implementation Details

As mentioned before, our approach does not depend on a specific method for fusion or deriving evidence about trustworthiness of information sources. However, in order to implement our approach, these methods need to be implemented as well. In this section, we reveal those implementations.

Each opinion corresponds to a set of evidence as described before. We fuse two or more opinions using the *consensus* (i.e., *cumulative fusion*) operator $\oplus$ of Subjective Logic [4]. Based on the consensus operator, the fusion of the opinions $\{w_y^0, \ldots, w_y^n\}$ is computed as in Equation 8, where $r(w)$, $s(w)$, and $op(r, s)$ are the functions defined in Section II.

$$fuse(w_y^0, \ldots, w_y^n) = op\left(\sum_{i=0}^{n} r(w_y^i), \sum_{i=0}^{n} s(w_y^i)\right) \tag{8}$$

This operator converts each opinion into corresponding positive and negative evidence using Equation 2, then sums all positive and negative evidences. Lastly, it converts the resulting evidence back to the corresponding opinion using Equation 1.

To derive some evidence about the trustworthiness of the information source i, the information consumer j compares its own opinions with the original opinions of i (i.e., i's opinions without any discounting) about the same propositions. In trust literature, expectation values of opinions are commonly used for decision making [7], and we follow the same practice and use expectation values to compare opinions. Let w_x^i and w_x^j be

opinions of i and j about the proposition x. Then, j can use these opinions to derive evidence about i's trustworthiness as an information source, as described below, where δ_p and δ_n are thresholds ($0 \leq \delta_p < 0.5 < \delta_n \leq 1$), and $\Delta(w_x^j, w_x^i)$ is the difference between the expectation values of these opinions. We set $\delta_p = 0.25$, $\delta_n = 0.75$, and $a_x^j = 0.5$ in our simulations.

$$\Delta(w_x^i, w_x^j) = |E(w_x^j, a_x^j) - E(w_x^i, a_x^j)|$$

- w_x^i is regarded as a positive evidence for i if $\Delta(w^i, w^\tau) \leq \delta_p$,
- w_x^i is regarded as a negative evidence for i if $\Delta(w^i, w^\tau) \geq \delta_n$.

While implementing TRIBE, we use 0.9 as the a priori trust for information sources. Therefore, if no evidence exists about the trustworthiness of a source, its trustworthiness is taken as 0.9 based on Equation 6. That is, in our simulations, TRIBE is optimistic about the behaviour of an information source as long as its opinions do not conflict with that of others. To solve conflicts, TRIBE composes a constraint optimisation problem; in order to find the optimum solution to the composed problem, we use constrained hill climbing [1].

B. Benchmarking Approaches

We compare our approach with four other approaches, which are described shortly below.

- **Consensus Only ($\oplus$):** Opinions are fused using the consensus operator without being discounted first using the trustworthiness of their sources.
- **Consensus with Discounting ($\otimes\oplus$):** Opinions provided by information sources are, first, discounted by their trustworthiness, then, the discounted opinions are fused using the consensus operator. To compute trustworthiness of the sources, Equation 5 is used, where the positive and negative evidence is derived as described in the previous section.
- **Beta Reputation System (BRS):** In this approach, information sources provide their ratings about the validity of a binary proposition x as (r_x, s_x) pairs, where r_x is the amount of positive evidence and s_x is the amount of negative evidence for x. These ratings correspond to subjective opinions as described in Section II. The provided ratings are fused using the consensus fusion operator (i.e., by computing total positive and negative evidence) as in Equation 8. Whitby *et al.* extended BRS to filter out unfair ratings (i.e., misleading opinions) provided by the information sources. This approach filters out those ratings that do not comply with the significant majority of the ratings by using an *iterated filtering approach* [11]. Hence, this approach assumes that the majority of sources honestly share their opinions; liars are in the minority.
- **TRAVOS:** This approach is proposed by Teacy *et al.* [10] and very similar to BRS. The main difference between BRS and TRAVOS is the way they filter misleading opinions (i.e., unfair ratings). While BRS uses the majority of ratings to filter out unfair ratings from information sources, TRAVOS uses the personal observations about these sources to derive some evidence about their trustworthiness. That is, TRAVOS keeps a history of information sources and their opinions about propositions. To measure the trustworthiness of a source, an information consumer counts how many times their opinions agree and disagree for the same propositions. The number of agreements and disagreements are taken as amounts of positive and negative evidence and used to model trust in the source using beta distributions. Hence, unlike BRS, TRAVOS does not assume that the majority of the provided opinions are trustworthy; however, it requires an opinion history of a source to estimate its trustworthiness. If there is no evidence to estimate trust in a source, TRAVOS takes 0.5 as its trustworthiness.

BRS and TRAVOS are well-known approaches in trust and reputation literature. They are originally proposed to fuse opinions of information sources about trust-related binary proposition like "Agent i is trustworthy". However, these approaches are flexible enough to fuse opinions about any binary propositions. Therefore, in this paper, we use them to evaluate our approach.

C. Simulations and Results

We have conducted simulations to test our approach in various settings. During a simulation, an information consumer makes decisions throughout 50 time steps. At each simulation, there is one information consumer who makes the decision, in each time step t, about a *new proposition* p^t. For this purpose, the consumer requests opinions about the given proposition from each information source in a society of 20 information sources. The ground truth about the proposition p^t is represented as an opinion, which is either $(0.99, 0.01)$ or $(0.01, 0.99)$, but it is hidden from the information consumer and sources. Each source i can observe only a set of evidences Φ_i^t about p^t; the number of evidences in this set is randomly chosen between 3 and 10, i.e., $3 \leq |\Phi_i^t| \leq 10$. Let the ground truth be (b^τ, d^τ), then an evidence $e \in \Phi_i^t$ is a positive evidence with probability b^τ and a negative evidence with probability d^τ. Based on the observed evidence, the sources compose their opinion about p^t using Equation 1. An honest source shares its opinion with the information consumer without any modification. However, if the source is not honest (i.e., liar), it negates its genuine opinion (by swapping *belief* and *disbelief*) before sharing it with the consumer. In this way, liars aim at providing misleading opinions to the information consumer.

After collecting opinions about p^t from the sources, the consumer fuse them using one of the five approaches[1] and uses the fused opinion for decision making. We compute the performance of the information consumer at time step t based on the *absolute error* between the fused opinion and the ground truth. Let w^f and w^τ be the fused opinion and the ground truth, respectively, at time t. Then, the absolute error of w^f is computed as the absolute value of the difference between its expectation value and the expectation value of w^τ:

$$error(w^f | w^\tau) = \Delta(w^\tau, w^f) = |E(w^\tau, 0.5) - E(w^f, 0.5)|$$

[1]These approaches are referred to as ($\oplus$), ($\otimes\oplus$), BRS, TRAVOS, and TRIBE in the figures showing our results.

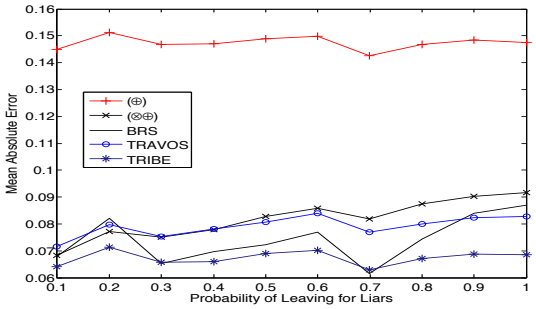

Fig. 2. Mean absolute error for $0.1 \leq P^{liar}_{leave} \leq 1$ when $R_{liar} = 0.1$.

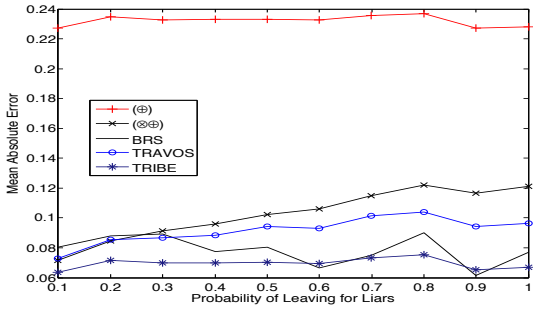

Fig. 3. Mean absolute error for $0.1 \leq P^{liar}_{leave} \leq 1$ when $R_{liar} = 0.2$.

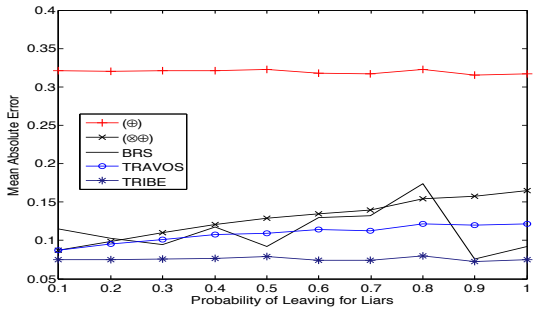

Fig. 4. Mean absolute error for $0.1 \leq P^{liar}_{leave} \leq 1$ when $R_{liar} = 0.3$.

At the end of each time step (i.e., after fusion and decision-making), the information consumer can also observe a set of evidences about the proposition p^t. The amount of observed evidence is also selected randomly between 3 and 10, as described before. Based on the evidence, the consumer generates its own opinion w^j about the proposition p^t and uses this opinion only to derive evidence about the trustworthiness of each information source i. For this purpose, it uses the technique described in the previous section (i.e., using $\Delta(w^j, w^i)$).

The ratio of liars among the information sources is determined by the parameter $R_{liar} \in \{0.1, 0.2, \ldots, 0.9\}$. To have dynamism in our simulations, we make honest sources leave the society with the probability 0.1 at the end of each time step. On the other hand, liars leave the society with a probability $P^{liar}_{leave} \in \{0.1, 0.2, \ldots, 0.9, 1\}$ at the end of each time step. When an information source leaves the society, a new information source of the same type joins to keep R_{liar} and the number of sources unchanged. Therefore, P^{liar}_{leave} allows us to simulate *whitewashing* attacks in Peer-to-Peer systems, where malicious agents leave the society when their reputation decrease and join back with new identities to whitewash their bad reputation and abuse the system.

For each pair of R_{liar} and P^{liar}_{leave} values, 10 simulations are conducted with different random number seeds. Therefore, 4500 simulations are run to test the five approaches with respect to different R_{liar} and P^{liar}_{leave} values. Here, we present only the average of our results, which are significant based on *t-test* with a confidence interval of 0.95.

Figure 2 shows our results when the ratio of liars is 0.1 and P^{liar}_{leave} is varied from 0.1 to 1. In this setting, the error is around 0.145 when only the consensus operator, i.e., $(\oplus)$, is used and it does not change much as P^{liar}_{leave} is varied. The error becomes 0.07 when $P^{liar}_{leave} = 0.1$ and increases to 0.09 as P^{liar}_{leave} increases if discounting is used before applying the consensus operator, i.e., $(\otimes\oplus)$. The error slightly lowers further if TRAVOS is used instead of $(\otimes\oplus)$. TRIBE provides the best performance, i.e., the lowest mean absolute error; the error does not go above 0.07 if TRIBE is used. The second best performance belongs to BRS. This is intuitive because the significant majority of information sources are honest in this setting, just as assumed by BRS.

Figure 3 shows our results when the ratio of liars is increased to 0.2. In this setting, we have similar results. The error increase to 0.23 when only consensus operator is used. Again, the best performance belongs to TRIBE; the error does not go above 0.07 when TRIBE is used. BRS has a similarly good performance with a mean absolute error around 0.075. The error is higher if $(\otimes\oplus)$ or TRAVOS is used, and slightly increases as P^{liar}_{leave} increases.

Figure 4 and Figure 5 show our results when the ratio of liars is increased to 0.3 and 0.4, respectively. In these figures, the error increases to 0.33 and 0.4, respectively, when only consensus operator is used. Similarly, the performances of $(\otimes\oplus)$, BRS, and TRAVOS decreases slightly in these settings. However, the highest performance is achieved again by TRIBE; the mean absolute error is around 0.08 when TRIBE is used.

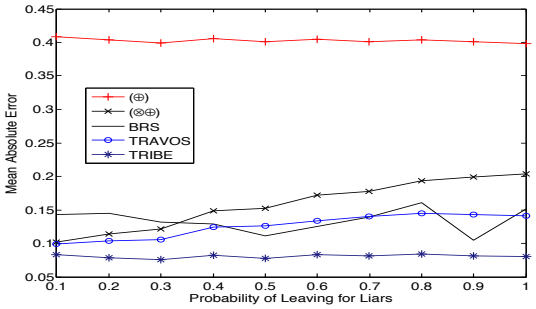

Fig. 5. Mean absolute error for $0.1 \leq P^{liar}_{leave} \leq 1$ when $R_{liar} = 0.4$.

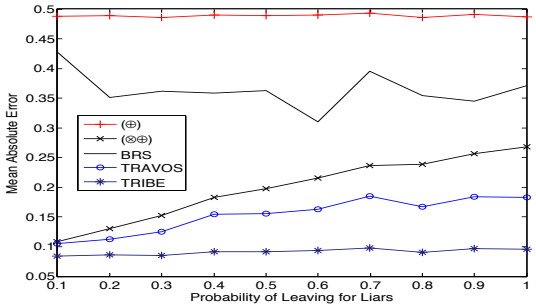

Fig. 6. Mean absolute error for $0.1 \leq P^{liar}_{leave} \leq 1$ when $R_{liar} = 0.5$.

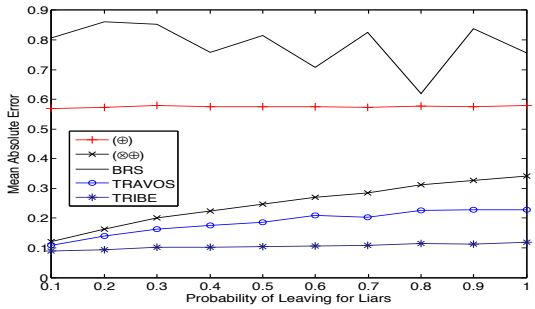

Fig. 7. Mean absolute error for $0.1 \leq P_{leave}^{liar} \leq 1$ when $R_{liar} = 0.6$.

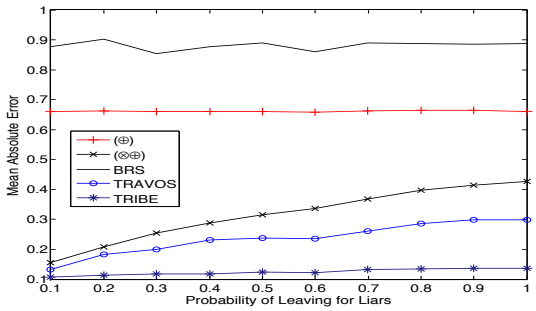

Fig. 8. Mean absolute error for $0.1 \leq P_{leave}^{liar} \leq 1$ when $R_{liar} = 0.7$.

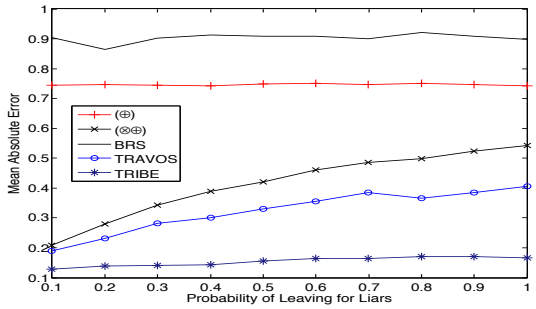

Fig. 9. Mean absolute error for $0.1 \leq P_{leave}^{liar} \leq 1$ when $R_{liar} = 0.8$.

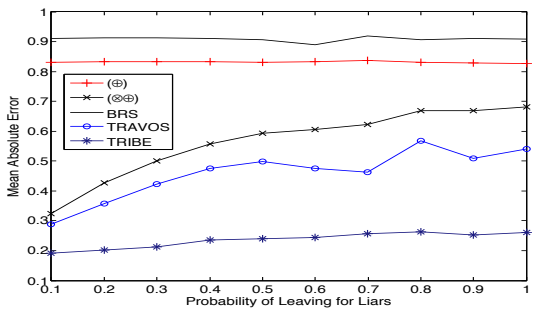

Fig. 10. Mean absolute error for $0.1 \leq P_{leave}^{liar} \leq 1$ when $R_{liar} = 0.9$.

Figure 6 shows our results when the ratio of liars is increased to 0.5. This is the setting where the number of malicious and honest sources are the same. The error in fusion when only consensus operator is used has increased to 0.5. In this setting, the error of BRS in fusion has dramatical increased and fluctuated between 0.33 and 0.43. This is because, the majority of sources are not honest any more. The mean absolute errors of $(\otimes\oplus)$ and TRAVOS are as low as 0.1 when $P_{leave}^{liar} = 0.1$; however, their errors increase to 0.27 and 0.17, respectively, as P_{leave}^{liar} increases. Unlike these approaches, the error of TRIBE in fusion is very low and around 0.08.

In our simulations, performances of $(\otimes\oplus)$ and TRAVOS decrease as P_{leave}^{liar} increases. This is intuitive because, as P_{leave}^{liar} increases, it gets harder to accumulate enough evidence about malicious sources to model their trustworthiness. Unlike these approaches, TRIBE can exploit conflicts between opinions as evidence to discount untrustworthy opinions. That is why TRIBE is robust to the variations in the probability that liars leave and rejoin the society.

Figure 7 and Figure 8 show results when the ratio of liars is increased to 0.6 and 0.7, respectively. In these settings, liars are the majority. Therefore, BRS considers opinions from malicious sources more trustworthy than the ones from honest sources. Hence, BRS has the worst performance; even using only consensus without discounting is better than BRS. High ratio of liars significantly increases the error of $(\otimes\oplus)$ and TRAVOS, while TRIBE has still very low error around 0.1.

Figure 9 and Figure 10 show our results when the ratio of liars is increased to 0.8 and 0.9, respectively. In these settings, significant majority of information sources are malicious. As a result, the error in fusion is dramatically high in these settings for all approaches except TRIBE. When $R_{liar} = 0.8$, the error in fusion is around 0.15 for TRIBE. The error of TRIBE slightly increases and resides between 0.2 and 0.25 when R_{liar} is increased to 0.9. This is an impressive performance when it is considered that only 10% of sources are honest and the remaining 90% are malicious. The success of TRIBE is based on detecting conflicts between opinions and discounting them further to resolve these conflicts. Although, the honest information sources are few in number, it is harder to discount their opinions, because more evidence is available about their trustworthiness due to their lower probability of leaving the society. Therefore, it is easy to discount opinions of malicious information sources, even though they are abundant.

V. Discussion

There are several models for computing trust and reputation in multiagent systems. In these models, direct evidence is combined with indirect evidence to model trust in agents. Direct evidence is based on personal observations, while indirect evidence is received from other agents that serve as information sources. Jøsang and Ismail proposed the beta reputation system (BRS) [6]. It estimates the likelihood of proposition "Agent i is trustworthy" – i.e., trustworthiness of the agent i – using beta probability density functions. For this purpose, aggregation of direct evidence and indirect evidence (i.e., ratings) from information sources are used as the parameters of beta distributions. Evidence shared by sources are equivalent to binary opinions in Subjective Logic [5]. Whitby *et al.* extended BRS to handle misleading opinions from malicious sources using a majority-based algorithm [11]. Teacy *et al.* proposed TRAVOS [10], which is similar to BRS, but it uses personal observations about information sources to estimate their trustworthiness as we do in this paper.

Yu and Singh proposed a trust approach that handles misleading indirect evidence using a version of weighted majority algorithm [14]. In their algorithm, weights are assigned to information sources. These weights are initiated as 1.0 and can be considered as the trustworthiness of the corresponding sources. The algorithm makes predictions about trust related

propositions (e.g, i is trustworthy) based on the weighted sum of indirect evidence (i.e., ratings) provided by those sources. The authors proposed to tune the weights after an unsuccessful prediction so that the weights assigned to the unreliable sources are decreased. They assume that the ratings from dishonest sources may conflict with the personal observations. By decreasing the weights of these sources over time, misleading ratings are filtered. Zhang and Cohen proposed the personalized approach that measures the trustworthiness of an information sources using two metrics [15]: 1) private reputation calculated by comparing the opinion of the source with the personal observations and 2) public reputation estimated by comparing the opinion of the source and the opinions from other sources about the same propositions.

In this paper, we describe conflicts between binomial opinions and propose an approach to resolve conflicts before performing fusion. Conflicts in knowledge lead to inconsistencies that hamper the reasoning over the knowledge. Therefore, before using such knowledge bases, their conflicts should be resolved. Gobeck and Halaschek [3] present a belief revision algorithm for ontologies, which is based on trust degrees of information sources to remove conflicting statements from a knowledge base. However, as the authors point out, the proposed algorithm is not guaranteed to be optimal. Dong *et al.* [2] propose to resolve conflicts in information from multiple sources by a voting mechanism. Double counting in votes is avoided by considering the information dependencies among sources. The dependences are derived from Bayesian analysis.

When considering multiple sources espousing multiple claims, it is possible to estimate their reliability through corroboration without direct and/or indirect evidence. For example, fact-finding algorithms aim to identify the *truth* given conflicting claims. Yin *et al.* proposed TruthFinder [13] which utilizes an iterative approach to estimate trustworthiness of information sources and information they provide. Their approach based on the assumption that a source is trustworthy if it provides many pieces of true information, and a piece of information is likely to be true if it is provided by many trustworthy sources. Therefore, very similar to BRS, TrustFinder also assumes that the information provided by the majority is trustworthy. In this work, we show that this assumption may lead to incorrect estimation of trust in information.

Ideally, information fusion creates a product that is better than what it would have been possible if the information pieces were taken individually. However, information fusion is a complex operation due to the uncertainties attached with the information – such as reliability, accuracy, and so forth of the information. There are many ways to fuse and reason about uncertain information. In this regard, *evidence theory* (i.e., DST) is a well-known mathematical framework to represent and fuse information with uncertainty. An important property to observe in DST is that information is assumed to be *independent*. There are numerous operators to fuse information – Dempster's rule, Yager's rule, and Inagaki's combination operator, to name a few [12].

To fuse opinions in our implementation of TRIBE, we use *consensus* operator of Subjective Logic [4]. The consensus operator provides a method for combining possibly conflicting beliefs within the Dempster-Shafer belief theory, and represents an alternative to the traditional Dempster's rule. In future, we plan to analyse the performance of TRIBE when it is used with other fusion operators.

In this work, we consider only binary frames, which are the propositions that take only two mutually exclusive values. Therefore, the opinions considered in this paper are binomial. In case the frame is larger than binary, then opinions are called multinomial, instead of binomial. We plan to extend our approach in future to accommodate multinomial opinions.

VI. Conclusions

In this paper, we propose TRIBE, which allows efficient identification of conflicting opinions. Then, these conflicts are resolved by trust revision using an approach based on constraint optimisation. Through simulations, we show that our approach can successfully handle highly misleading information in challenging settings. The simulations also show that the approach is robust in the face of liars that whitewash bad evidence about their trustworthiness by leaving and re-joining the society. In this paper, we study only binomial opinions about binary propositions. In the future, we plan to extend our approach to handle multinomial opinions and evaluate its performance when used with different fusion operators.

References

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[2] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: The role of source dependence. In *Proc. of the 35th International Conference on Very Large Databases*, Lyon, France, August 2009.

[3] J. Golbeck and C. Halaschek-Wiener. Trust-based revision for expressive web syndication. *Journal of Logic and Computation*, 19(5):771–790, Oct. 2009.

[4] A. Jøsang. The consensus operator for combining beliefs. *Artificial Intelligence Journal*, 142:157–170, 2002.

[5] A. Jøsang. *Subjective Logic*. Book Draft, 2011.

[6] A. Jøsang and R. Ismail. The beta reputation system. In *Proc. of the 15th Bled Electronic Commerce Conference e-Reality: Constructing the e-Economy*, pages 48–64, 2002.

[7] A. Jøsang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decis. Support Syst.*, 43:618–644, March 2007.

[8] A. Preece, D. Pizzocaro, D. Braines, D. Mott, G. de Mel, and T. Pham. Integrating hard and soft information sources for D2D using controlled natural language. In *Proceedings of the International Conference on Information Fusion*, pages 1330–1337, 2012.

[9] G. Shafer. *A mathematical theory of evidence*. Princeton university press, 1976.

[10] W. Teacy, J. Patel, N. Jennings, and M. Luck. TRAVOS: Trust and reputation in the context of inaccurate information sources. *Autonomous Agents and Multi-Agent Systems*, 12(2):183–198, 2006.

[11] A. Whitby, A. Jøsang, and J. Indulska. Filtering out unfair ratings in Bayesian reputation systems. *The Icfain Journal of Management Research*, 4(2):48–64, 2005.

[12] R. Yager, editor. *Classic works of the Dempster-Shafer theory of belief functions*. Springer, 2008.

[13] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. In *Proceedings of the Conference on Knowledge and Data Discovery*, 2007.

[14] B. Yu and M. Singh. Detecting deception in reputation management. In *Proceedings of Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 73–80, 2003.

[15] J. Zhang and R. Cohen. A comprehensive approach for sharing semantic web trust ratings. *Computational Intelligence*, 23(3):302–319, 2007.