Reasoning about Uncertain Information and Conflict Resolution through Trust Revision

Murat Şensoy¹,⁴, Achille Fokoue², Jeff Z. Pan³, Timothy J. Norman¹, Yuying Tang¹, Nir Oren¹, and Katia Sycara³
¹Computing Science, University of Aberdeen, UK
²IBM T. J. Watson Research Center, NY, US
³Carnegie Mellon University, Pittsburgh, US
⁴Computer Science, Ozyegin University, Istanbul, Turkey
{m.sensoy,jeff.z.pan,t.j.norman,n.oren}@abdn.ac.uk,
achille@us.ibm.com, {yuying.tang,katia}@cs.cmu.edu

ABSTRACT
In information driven MAS, information consumers collect information about their environment from various sources such as sensors. However, there is no guarantee that a source will provide the requested information truthfully and correctly. Even if information is provided only by trustworthy sources, it can contain conflicts that hamper its usability. In this paper, we propose to exploit such conflicts to revise trust in information. This requires a reasoning mechanism that can accommodate domain constraints, uncertainty, and trust. Our formalism — SDL-Lite — is an extension of a tractable subset of Description Logics with Dempster-Shafer theory of evidence. SDL-Lite allows reasoning about uncertain information and enables conflict detection. Then, we propose methods for conflict resolution through trust revision and analyse them through simulations. We show that the proposed methods allow reasonably accurate estimations of trust in information in realistic settings.

Categories and Subject Descriptors
I.2.11 [Distributed Artificial Intelligence]: Multiagent Systems

Keywords
Information Fusion, Trust, Uncertainty, Description Logics

1. INTRODUCTION
Uncertainty is a core feature of many domains in which agents are expected to operate. In such environments, information sources such as sensors provide agents with state information. However, in the context of a multi-agent system, different sets of sensors can be controlled by different agents, each with their own capabilities. In order to reason about the state of the environment, an agent must therefore request potentially noisy, incomplete, or misleading information from potentially untrustworthy agents. By obtaining information from multiple sources, the agent can build up a more accurate view of its environment than by utilising its own sensors alone. To achieve such a task, the agent needs to reason about the noisy, incomplete, and misleading information. This paper addresses such a need with a mechanism that can reason about uncertain information and conflict resolution through trust revision.

Various researchers have examined aspects of this problem — work in computational trust (e.g. [10]) is intended to allow an agent to determine which other agents should be asked for information, while work on information fusion [13] examines how incomplete and noisy information from different sources should be combined in order to obtain a true picture of the environment. However, neither work in isolation considers both sources of uncertainty in the domain, namely trustworthiness of information sources as well as incompleteness and vagueness of the provided information.

This paper makes two core contributions: Firstly, we combine Description Logics (DLs) [1] and Dempster-Shafer theory (DST) [15] to create a computationally tractable framework for reasoning about uncertain information obtained from different agents. This framework enables us to detect conflicts in uncertain information due to constraints imposed by the domain. It also enables us to resolve such conflicts. Secondly, we show how trust in uncertain information can be revised when additional information is received. We model this problem as an optimisation problem, and propose heuristics that allow us to identify high quality solutions.

The remainder of the paper is structured as follows. We begin by introducing Description Logics and Dempster-Shafer theory. Section 3 describes SDL-Lite, our extension of DL-Lite with subjective opinions. In Section 4, we concentrate on the problem of how much trust should be placed in a source of information, and we describe how conflicting information sources can be made consistent. We evaluate our approach in Section 5 and discuss it with respect to the existing work and future research directions in Section 6.

2. PRELIMINARIES

2.1 DL-based Ontologies
Due to limited space, we do not provide a full overview of Description Logics (DLs), but rather point the reader to [1]. We note,
however, that even for the smallest propositionally closed DL, $\mathcal{AC}^C$ (which only provides class constructors $\neg C, C \cap D, C \cup D, \exists R.C$ and $\forall R.C$), the complexity of entailment is EXPTIME. Recently, Calvanese et al. [5] proposed DL-Lite, which can express most features in UML class diagrams with a low reasoning overhead (with data complexity $AC^O$). It is for this reason that we base our model on DL-Lite-core (referred to here as DL-Lite, although there are extensions [3]), and hence provide a brief formalisation to ground the subsequent presentation of our model.

A DL-Lite knowledge base $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ consists of a TBox $\mathcal{T}$ and an ABox $\mathcal{A}$. Axioms of the forms consist of $K$:

1. class inclusion axioms: $B \sqsubseteq C$ in $\mathcal{T}$ where $B$ is a basic class $B := A \mid \exists R \mid \exists R^\bot$ and $C$ is a general class $C := B \mid \neg B \mid C_1 \sqcap C_2$ (where $A$ denotes an named class, $R$ denotes a named property, and $R^\bot$ is the inverse of $R$);

2. individual axioms: $B(a), R(a, b) \in \mathcal{A}$ where $a$ and $b$ are named individuals.

Description Logics have a well-defined model-theoretic semantics, which are provided in terms of interpretations. An interpretation $\mathcal{I}$ is a pair $(\Delta^I, \mathcal{M})$, where $\Delta^I$ is a non-empty set of objects and $\mathcal{M}$ is an interpretation function, which maps each class $C$ to a subset $C^I \subseteq \Delta^I$ and each property $R$ to a subset $R^I \subseteq \Delta^I \times \Delta^I$.

Using a trivial normalisation, it is possible to convert class inclusion axioms of the form $B_1 \sqsubseteq C \sqcap C_2$ into a set of simpler class inclusions of the form $B_1 \sqsubseteq C_3$ or $B_3 \sqsubseteq -B_1$, where $B_1$, $B_3$, and $B_2$ are basic concepts [5]. For instance, during normalisation, $B_1 \sqsubseteq B_2 \sqcap -B_3$ is replaced with $B_1 \sqsubseteq B_2$ and $B_1 \sqsubseteq -B_3$. In Table 3, we define semantics over a normalised TBox for our variant of DL-Lite — SDL-Lite.

### 2.2 Subjective Opinions

Dempster-Shafer Theory (DST) offers a 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 DST, a binomial opinion about a proposition $x$ is represented by a triple $w_x = (b_x, d_x, u_x)$ which is derived from the basic probability masses assigned to subsets of truth assignments of the language. 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 $0 \leq b_x, d_x, u_x \leq 1$. When a more concise notation is necessary, we use $(b_x, d_x)$ instead of $(b_x, d_x, u_x)$ since $u_x = 1 - b_x - d_x$. The negation over an opinion $w_x$ is defined as $\neg w_x = (b_x, d_x, u_x) = (d_x, b_x, u_x) = (b_x, d_x, u_x)$ [9].

**Definition 1.** Let $w_1 = (b_1, d_1, u_1)$ and $w_2 = (b_2, d_2, u_2)$ be two opinions about the same proposition. We call $w_1$ a specialisation of $w_2 (w_1 \leq w_2)$ if $b_2 \leq b_1$ and $d_2 \leq d_1$ (implies $u_1 \leq u_2$). Similarly, we call $w_1$ a generalisation of $w_2 (w_2 \leq w_1)$ if $b_1 \leq b_2$ and $d_1 \leq d_2$ (implies $u_2 \leq u_1$).

An agent’s opinion about a proposition $x$ is denoted by $w_x = (b_x, d_x, u_x)$. This opinion $w_x$ 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 [15] proposed a discounting operator $\otimes$ to normalise the belief and disbelief in $w_x$ based on the degree of trust $j$ has of $i$ with respect to $x$: $t^j_i$. The normalised opinion, $w^j_x$, is computed as $(b^j_x \times t^j_i, d^j_x \times t^j_i)$. The trustworthiness of information sources can be modelled using Beta probability density functions [10]. A Beta distribution has two parameters $(r + 1, s + 1)$, where $r$ is the amount of positive evidence and $s$ is the amount of negative evidence for the trustworthiness agent $i$ agent has for agent $j$. The degree of trust $t^j_i$ is then computed as the expectation value of the Beta distribution: $t^j_i = (r + 1)/(r + s + 2)$.

We now provide a running example used throughout the remainder of the paper. Following this, we will present the semantics and reasoning mechanisms for a language that combines DL-Lite and DST with the view to offering a model for tractable reasoning with uncertain information and trust.

### 2.3 Example Scenario

Consider a region in which insurgents are active and where civilian groups are in need of support. An NGO operating in the region has identified a safe zone Z and aims to bring relief to the injured in village V by transporting them to the safe zone. There is only one road between Z and V, but there is conflict between groups $G_1$ and $G_2$ within the region. As part of a multi-national peace effort, $M_1$ and $M_2$ operate within the region, and part of their remit is to protect and support NGOs. The resources available to $M_2$ include Unmanned Aerial Vehicles (UAVs). $M_1$ acts as liaison to the NGO, and has intelligence from the information sources listed in Table 1 along with models of the trustworthiness of these sources. $M_1$ collects the following pieces of information from the sources in the area: 1) $M_2$ informs $M_1$ that $R$ is blocked with opinion $(0.8, 0.1, 0.1)$; 2) $P$ reports that $R$ is safe with opinion $(0.8, 0.1, 0.1)$; and 3) $A$ reports an explosion on route $R$ with opinion $(0.2, 0.3, 0.5)$. This intelligence is interpreted by $M_1$ given its degree of trust in the sources into Table 1. The opinions from information sources are discounted by the trustworthiness of the sources. For instance, $M_2$’s original opinion for ‘$R$ is blocked’ is $(0.8, 0.1, 0.1)$, but it is discounted to $(0.71, 0.09, 0.2)$ using $M_2$’s trustworthiness, i.e., 0.89 (see Table 2 Abox for other discounted opinions).

Having provided an overview of DLs and DST, we now turn to describing our core contribution — a description logic able to represent uncertainty in the manner of DST.

### 3. SUBJECIVE DL-LITE

We propose *Subjective DL-Lite* (or SDL-Lite for short), which extends DL-Lite-core with subjective opinion assertions of the form $B \triangleright w$, where $w$ is an opinion and $B$ is an ABox axiom (i.e., assertion). Each ABox axiom is associated with one opinion. ABox axioms have the form $B(a) \triangleright R(a, b)$, where $B$ is basic class, $R$ is a property, and $a$ and $b$ are individuals.

#### 3.1 SDL-Lite Semantics

In common with DL-Lite ontologies, the semantics of an ontology in SDL-Lite is defined in terms of subjective interpretations. Let $W$ be the set of all possible subjective binary opinions, a subjective interpretation is a pair $\mathcal{I} = (\Delta^I, \mathcal{M})$ where the domain $\Delta^I$...
Table 3: Semantics of Subjective DL-Lite

<table>
<thead>
<tr>
<th>Syntax</th>
<th>Semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T ⊲ (a, 0, 0)</td>
</tr>
<tr>
<td>⊥</td>
<td>⊥ ⊲ (a, 0, 1)</td>
</tr>
<tr>
<td>∃R b ⊲ (Rb ⊲ a1) ≥ max ⊲ {b ⊲ (Rb ⊲ a2)} and</td>
<td></td>
</tr>
<tr>
<td>d ⊲ (Rb ⊲ a1) ≤ min ⊲ {d ⊲ (Rb ⊲ a2)} and</td>
<td></td>
</tr>
<tr>
<td>¬B · B ≤ · B</td>
<td>¬B ⊲ (a, 0)</td>
</tr>
<tr>
<td>R⁻</td>
<td>R⁻ ⊲ (a1, 0, 1) = R⁻ ⊲ (a2, 0, 2)</td>
</tr>
</tbody>
</table>

B₁ ⊲ B₂ ∀a ∈ ∆², b(B₁ ⊲ a) ≤ b(B₂ ⊲ a) and
| d(B₁ ⊲ a) ≤ d(B₂ ⊲ a)                                    |
B₁ ⊲ ¬B₂ ∀a ∈ ∆², b(B₁ ⊲ a) ≤ d(B₂ ⊲ a) and
| b(B₁ ⊲ a) ≤ d(B₂ ⊲ a)                                    |

B(a) ⊲ w b(w) ≤ b(B ⊲ a) and d(w) ≤ d(B ⊲ a)
| R(a, b) ⊲ w b(w) ≤ b(R ⊲ a, b) and d(w) ≤ d(R ⊲ a, b) |

is a non-empty set of objects and ∆² is a subjective interpretation function, which maps:

- an individual a to an element of a ⊲ ∆²;
- a named class A to a function A ⊲ : ∆² → W;
- a named property R to a function R² : ∆² × ∆² → W.

To provide a semantics for SDL-Lite, we extend interpretations of DL-Lite class and property descriptions, and of axioms under unique name assumption. The semantics are presented in Table 3.

The semantics of ∃R is derived from the rule R ⊲ ⊲ a ⊲ b → ∃R(α ⊲ b), ∀b ∈ ∆². This rule constrains the minimum belief and the maximum disbelief that R ⊲ a can have. For any individuals a and b, the belief in a having a property R (i.e., ∃R(a)), is not less than belief in a having the property R with b (i.e., R(a, b)), and disbelief in ∃R(a) is not more than disbelief in R(a, b).

An ontology provides us with domain constraints in the form of TBox axioms. For instance, the axiom B₁ ⊲ B₂ means that every instance of class B₁ is also an instance of class B₂. This trivially implies ¬B₂ ⊲ ¬B₁, i.e., an individual that is not an instance of B₁ cannot be an instance of B₂. Therefore, given an individual a, the axiom B₁ ⊲ B₂ implies that our belief in B₂(a) cannot be less than our belief in B₁(a) and our disbelief in B₂(a) cannot be more than our disbelief in B₁(a). That is, b(B₂ ⊲ a) ≤ b(B₁ ⊲ a) and d(B₂ ⊲ a) ≤ d(B₁ ⊲ a) must hold. Similar constraints also exist in Table 3 for B₁ ⊲ B₂.

Definition 2. An SDL-Lite knowledge base K = (T, A) is consistent if and only if it has a model. A model of K is an interpretation of K that satisfies the constraints in Table 3.

If K is consistent, it can have many models, but one of them is the most general model with respect to the partial ordering on opinions by Definition 1. In the next section, we describe how to detect consistency, and how to compute the most general model of a consistent SDL-Lite knowledge base.

3.2 Reasoning with SDL-Lite

The aims of this section are twofold. First, we detail the reasoning mechanisms necessary to infer opinions about the world given an SDL-Lite ontology from the information available to an agent. Second, we define the conditions under which an SDL-Lite knowledge base becomes inconsistent, and prove that our reasoning mechanism ensures the maintenance of a consistent knowledge base under certain conditions.

The TBox T of an SDL-Lite knowledge base contains: (i) positive inclusions (PIs) in the form B₁ ⊲ B₂, where B₁ and B₂ are basic concepts (being either an atomic or an existential concept); and (ii) negative inclusions (NIs) of the form B₁ ⊳ ¬B₂. In order to reason over an SDL-Lite knowledge base K = (T, A), we compute the extended knowledge base K* = (T*, A*).

The extended TBox T* is initialised as equivalent to T and expanded by computing all (nontrivial) NIs and PIs between basic concepts implied by T*; i.e. T* is closed with respect to the rules:

- if B₁ ⊳ B₂ occurs in T* and either B₂ ⊳ ¬B₃ or B₃ ⊳ ¬B₂ occurs in T*, then B₁ ⊳ ¬B₃ to T*.

We, therefore, extend T in O(n²) steps in the size of T and form T* that contains all possible inclusions derived from T.

The extended ABox A* is initialised as equivalent to A and expanded in O(n) steps in the size of A using the rules:

- if R(a, b) : w in A*, add R⁻ (b, a) : w.

We may now specify how interpretations of classes and roles may be computed based on the computed T* and A*. Given a class Bₙ, T* may contain the following axioms that constrain the interpretation of an ABox assertion Bₙ(a):

- B₀ ⊳ Bₙ, B₁ ⊳ Bₙ, ..., Bₙ ⊳ B₀
- Bₙ+1 ⊳ ¬Bₙ, Bₙ+2 ⊳ ¬Bₙ, ..., Bₙ ⊳ ¬B₀
- Bₙ ⊳ ¬Bₙ+1, Bₙ ⊳ ¬Bₙ+2, ..., Bₙ ⊳ ¬Bₙ

Given these kinds of axioms, our aim here is to define appropriate inferences regarding the opinion of some assertion in A*.

Let wⁿ refer to the opinion related to the ABox assertion Bₙ(a), i.e., Bₙ(a) : wⁿ ∈ A*. Clearly, if Bₙ(a) does not appear in A*—representing the case that we have no evidence regarding Bₙ(a) — then wⁿ = (0, 0, 1), representing maximal uncertainty regarding Bₙ(a). We use wⁿ to refer to the interpretation of Bₙ(a), i.e., wⁿ = Bₙ(a). The most general opinion wⁿ that satisfy the constraints for wⁿ is then computed as follows:

wⁿ = (max(S₁), max(S₂)) where
S₁ = {w¹, w₀, ...} and S₂ = {w₀, w⁻₁, ...}

(1) d(w⁺) + d(w⁻) > 1, then there is no opinion satisfying the constraints defined by the semantics for Bₙ(a); otherwise, we take wⁿ as an interpretation of Bₙ(a).

Let us explain the interpretations for class assertions through the scenario in Section 2.3. Table 4 shows the extended TBox for the scenario. The interpretation of SubtagedRoad = (B₁ ⊲ B₂) is constrained by the TBox axioms b₁, b₂, b₃, b₄, b₅, and t₁₀ of Table 4. Therefore, given the ABox in Table 2 and based on Equation 1, the interpretation is computed as (max(0, 0.3)) = (0.0, 0.63, 0.37).

The interpretation for R₁(a, b) is constrained by the interpretations of ∃R₁(a) and ∃R₂(b). Let R₁(a, b) : w ∈ A and ∃R₁(a) : w₁

839
be the ABox assertion added to $A'$ while extending $A$. We know that $b(w_m) \geq b(w)$ and $d(w) \geq d(w_m)$. Belief in $\exists R^+_n(a)$ does not constrain belief in $R^+_n(a, b)$. However, in a consistent knowledge base, disbelief in $R^+_n(a, b)$ cannot be lower than disbelief in $\exists R^+_n(a)$ or $\exists R^-_n(b)$. Hence, disbelief in $\exists R^+_n(a)$ is constrained by the following TBox axioms in $T^+$:

1. $B_i \notin \exists R_n, B_j \notin \exists R_n, \ldots, B_{i-1} \notin \exists R_n, B_i \notin \exists R_n$.
2. $\exists R_n \ni B_{i+1}, \exists R_n \ni B_{i+2}, \ldots, \exists R_n \ni B_{i-n}, \exists R_n \ni B_{i}$.
3. $\exists R_n \ni \neg B_{i+1}, \exists R_n \ni \neg B_{i+2}, \ldots, \exists R_n \ni \neg B_{i-n}, \exists R_n \ni \neg B_{i-n}$.

Disbelief in $\exists R^+_n(b)$ is constrained by the axioms in $T^+$:

1. $B_{i+1} \notin \exists R_n, B_{i+2} \notin \exists R_n, \ldots, B_{i-n} \notin \exists R_n, B_i \notin \exists R_n$.
2. $\exists R_n \ni B_{i+1}, \exists R_n \ni B_{i+2}, \ldots, \exists R_n \ni B_{i-n}, \exists R_n \ni \neg B_{i}$.
3. $\exists R_n \ni \neg B_{i+1}, \exists R_n \ni \neg B_{i+2}, \ldots, \exists R_n \ni \neg B_{i-n}, \exists R_n \ni \neg B_{i-n}$.

Let $w_i$ refer to the opinion related to the ABox assertion for $B_i(a)$ if $1 \leq i \leq c$ and for $B_i(b)$ if $c+1 \leq i \leq f$. The most general opinion $w^*$ that satisfies the constraints for $R^+_n(a, b)$ is computed as follows:

$$w^* = (b(w), \max(S_d))$$

where

$$S_d = \{d(w), b(w_1), \ldots, b(w_m), b(w_{c+1}), \ldots, b(w_d),$$

$$d(w_{c+1}), d(w_{c+2}), \ldots, d(w_n),$$

$$d(w_{d+1}), d(w_{d+2}), \ldots, d(w_e),$$

$$d(w_{e+1}), d(w_{e+2}), \ldots, d(w_f),$$

$$b(w_{e+1}), b(w_{e+2}), \ldots, b(w_f),$$

$$b(w_{f+1}), b(w_{f+2}), \ldots, b(w_g)\}.$$  (2)

If $b(w^*) + d(w^*) > 1$, there is no opinion satisfying the constraints defined by the semantics for $R^+_n(a, b)$; otherwise, we take $w^*$ as interpretation of $R^+_n(a, b)$.

Let us now compute the interpretation of $\text{roadBombedBy}(r, g_i)$ in our example scenario. It is constrained by the interpretations of two other assertions $\exists \text{roadBombedBy}(r)$ and $\exists \text{roadBombedBy}(g_i)$. The disbelief in the interpretation of $\text{roadBombedBy}(r)$ is constrained by TBox axioms $t_6, t_7, t_8, t_9$, and $t_{12}$ of Table 4. We have opinion assertions only for the assertions $\text{Blocked}(r)$, $\text{Safe}(r)$, and $\text{BombedRoad}(r)$ in the extended ABox derived from Table 2. Therefore, based on Equation 2, we compute the interpretation as $(0, \max(0, 0.63, 0.3, 0.09)) = (0, 0.63)$. Table 4 shows the interpretations computed for the scenario using Equations 1 and 2.

The computational complexity of these calculations is $O(n)$ in the size of $T^+$ and $A^+$. Now, we introduce Theorem 1, which defines the conditions necessary and sufficient for inconsistency.

**Theorem 1.** An extended SDL-Lite KB $K^* = (T^+, A^+)$ with a coherent $T^+$ is inconsistent with respect to the semantics in Table 3 if and only if one of the following conditions hold:

1. Given $B_m(a); w_m, B_n(a); w_n \in A^+$, and $B_n \in T^+$, we have $b(w_m) + d(w_m) > 1$.
2. Given $B_m(a); w_m, B_n(a); w_n \in A^+$, and $B_m \in \neg B_n \in T^+$, we have $b(w_m) + d(w_m) > 1$.

**Proof:** The inconsistency arises if and only if at least one class or role does not have a valid interpretation satisfying the semantics in Table 3. Let us first analyse the inconsistencies due to the interpretations of classes. The most general interpretation for $B_i(a)$ is computed as in Equation 1 and referred to as $w_i$. Let $w_i$ be the opinion for $B_i(a)$ in $A^+$. If $b(w_i^*) + d(w_i^*) > 1$, there is no opinion satisfying the constraints defined by the semantics for $B_i(a)$ and $K^*$ is inconsistent. To have $b(w_i^*) + d(w_i^*) > 1$, one of the following conditions must hold based on Equation 1:

- $d(w_i^*) \in \{b(w_{i+1}), b(w_{i+2}), \ldots, b(w_i)\}$: This implies that there exists a TBox axiom $B_m \in \neg B_n$ with ABox assertions $B_m(a); w_m$ and $B_n(a); w_n$ such that $b(w_m) + b(w_n) > 1$.
- $d(w_i^*) \in \{d(w_{i+1}), d(w_{i+2}), \ldots, d(w_i)\}$: This implies that there exists a TBox axiom $B_m \in \neg B_n$ with ABox assertions $B_m(a); w_m$ and $B_n(a); w_n$ such that $b(w_m) + d(w_n) > 1$.
- $d(w_i^*) \in \{b(w_{i+1}), b(w_{i+2}), \ldots, b(w_i)\}$: This implies that there exists a TBox axiom $B_m \in \neg B_n$ with ABox assertions $B_m(a); w_m$ and $B_n(a); w_n$ such that $b(w_m) + b(w_n) > 1$.

Now, we look into the interpretations for role axioms to test consistency. The most general interpretation for $R^+_n(a, b)$ is computed as in Equation 2 and referred to as $w^*$. Let $w$ be the opinion for $R^+_n(a, b)$ in $A^+$. If $b(w^*) + d(w^*) > 1$, there is no opinion satisfying the constraints defined by the semantics for $R^+_n(a, b)$ and $K^*$ is inconsistent. To have $b(w^*) + d(w^*) > 1$, one of the following conditions must hold based on Equation 2:

- $b(w_n^*) \in \{b(w_{n+1}), b(w_{n+2}), \ldots, b(w_n)\}$: This implies that there exists a TBox axiom $B_m \in \neg B_n$ with ABox assertions $B_m(a); w_m$ and $B_n(a); w_n$ such that $b(w_m) + b(w_n) > 1$.
- $b(w_n^*) \in \{d(w_{n+1}), d(w_{n+2}), \ldots, d(w_n)\}$: This implies that there exists a TBox axiom $B_m \in \neg B_n$ with ABox assertions $B_m(a); w_m$ and $B_n(a); w_n$ such that $b(w_m) + d(w_n) > 1$.
- $b(w_n^*) \in \{b(w_{n+1}), b(w_{n+2}), \ldots, b(w_n)\}$: This implies that there exists a TBox axiom $B_m \in \neg B_n$ with ABox assertions $B_m(a); w_m$ and $B_n(a); w_n$ such that $b(w_m) + b(w_n) > 1$.

As shown, in the case of inconsistency, one of the conditions defined in Theorem 1 must hold. Furthermore, if none of these conditions holds in $K^*$, we guarantee that $K^*$ is consistent.

In an inconsistent extended SDL-Lite knowledge base $K^* = (T^+, A^+)$, the inconsistencies exist only because of conflicting opinions. Two opinions $w_m$ and $w_n$, which are about $B_m(a)$ and $B_n(a)$ respectively, are in conflict if they satisfy one of the conditions in Theorem 1. We label the portion of $w_m$ which conflicts with $w_n$ as $c_m$, and refer to it as the conflicting portion. If the conflict is due to the axiom $B_m \in \neg B_n \in T^+$, then the conflict arises because $b(w_m) + d(w_n) > 1$; hence $c_m = b(w_m)$ and $c_n = d(w_n)$. On the other hand, if the conflict is due to the axiom $B_m \in \neg B_n \in T^+$, we have conflict because $b(w_m) + b(w_n) > 1$; hence $c_m = b(w_m)$ and $c_n = b(w_n)$. If all conflicts in $K^*$ are resolved, then the knowledge base becomes consistent.

In the rest of the paper, we assume that the opinion about a specific ABox assertions is provided by a single source. When there is more than one source for an assertion, only one of them is chosen (e.g., based on their trustworthiness). This will be relaxed in future.
Having described sDL-Lite we now examine a novel application of the system, describing how evidence from multiple sources can be reasoned about based on the trust placed in these sources.

4. TRUST-BASED EVIDENCE ANALYSIS

Here we get to the crux of the problem being addressed in this paper: how can we draw reliable conclusions regarding the state of the world, given evidence acquired from disparate sources (agents), about whom we have variable trust? We refer to this process as trust-based evidence analysis. Our aim is not to offer a new mechanism for assessing the trustworthiness of information sources; in fact, we exploit a widely-studied model [10] for this purpose based on Beta distributions as described in Section 2.2. The novelty of this work lies in the use of such models to guide evidence analysis.

4.1 Handling Inconsistencies

sDL-Lite presented in the previous section provides a tractable means to capture and interpret evidence acquired from other agents. The fact that we have evidence from multiple agents, however, means that there are likely to be inconsistencies in the evidence received. Thus, given evidence (i.e., opinions) from various sources, our knowledge-base may not be consistent. This is despite the use of discounting through DST. Discounting provides us with a “best-guess” of the reliability of agents based on an aggregation of our prior experiences with, and other knowledge of them as evidence sources. As with any computational model of trust, the trust assessments that drive discounting are vulnerable to: lack of evidence about other agents and the effects of whitewashing [2]; a conflation of the probability of malicious behaviour and lack competence/expertise in the evidence-provider; strategic liars; and collusion among evidence-providers. In our running example, for instance, local police and civilian sources have relatively low trustworthiness, not because of any perceived malicious intent but due to a belief that they lack experience in providing precise information. With more evidence, trustworthiness of information sources may be modelled more accurately, but our challenge is to support the analysis of evidence given the status quo.

To illustrate this challenge, consider an adaptation of our example (case II) in which additional evidence is received from a third source, agent C, about R: C reports that R was bombed by G1 with opinion (0.8, 0.1, 0.1). With this additional report, our ABox contains roadBombedBy(R, G1): (0.67, 0.083, 0.247) after discounting the opinion with C’s trustworthiness 0.83 listed in Table 1 (a3 in Table 5 where the resulting extended ABox is presented). The extended ABox will now have a conflict between a1 and a0, because 0.67 + 0.63 > 1 and the extended TBox contains roadBombedBy ⊢ ¬Safe. Let w1 = (0.63, 0.066, 0.304) and w2 = (0.67, 0.0, 0.33). The conflicting portions of w1 and w2 are c12 = 0.63 and c21 = 0.67. Let us refer to the trustworthiness of the sources of w1 and w2 as t1 and t2 respectively. In our example, from Table 1, t1 = 0.83 and t2 = 0.786. In order for us to transform our inconsistent knowledge-base into a consistent knowledge-base, from which we can draw valid conclusions given our semantics, we need to determine additional discounting factors x1 and x2 for opinions w1 and w2 such that 0 ≤ c12.x1 + c21.x2 ≤ 1.

In this paper, we specify this problem as that of finding additional discounting factors for the belief-mass distributions of pieces of evidence to make our knowledge-base consistent. In general, our conflict resolution problem is a tuple (C, X) where C is the set of conflicting portions that appear in the extended knowledge base K∗, and X is a set of additional discounting factors corresponding to C. We require that, in (C, X), ∀cij ∈ C, 3cij ∈ C and 3xj ∈ X. Then, a solution to this problem is an assignment of values to each xj ∈ X such that

\[ \forall c_{ij}, c_{ij} \in C, \forall x_i, x_j \in X \quad 0 \leq c_{ij}.x_i + c_{ij}.x_j \leq 1 \]

There are many heuristic approaches to solving this problem, among them being to consider only consistent knowledge to draw conclusions from the evidence received; i.e. ∀x_i ∈ X, x_i = 0. This, however, could lead to a significant loss of evidence. Here, we explore a nuance of more refined approaches that guarantee the generation of a consistent knowledge-base: trust-based deleting, trust-based discounting and evidence-based discounting.

4.2 Trust-based deleting

If two opinions w1 and w2 are in conflict, the opinion from the less trustworthy source is deleted, and if both sources are equally trustworthy both opinions are deleted. Thus, if the trust we have in the source of opinion w1 is greater than that of the source of w2 (t1 > t2) then x2 = 0 and x1 = 1, and in the event that t1 = t2 we assign x1 = x2 = 0. In our example, the local police sources P are slightly less trustworthy than the local civilian sources C. Hence, the opinion about Safe(R) is changed to (0, 0.1) and the conflict is resolved. This approach, however, neglects the amount of evidence used to calculate trust, and it does not consider the difference between trust values (tC = 0.83 ≈ tP = 0.786).

4.3 Trust-based discounting

If two opinions w1 and w2 are in conflict, they are discounted in proportion to the trustworthiness of their sources. That is, the additional discounting factor for w1 and w2 is computed using t1/(c12.t1 + c21.t2) and t2/(c12.t1 + c21.t2), respectively, where t1 and t2 are the trustworthiness of the sources of the opinions. In our example, an additional discount factor of roadBombedBy(R, G1) is 0.79 and that of Safe(R) is 0.75, since the trustworthiness of C and P are 0.83 and 0.786, respectively. Therefore, to resolve the conflict, the original opinion of C about roadBombedBy(R, G1) is discounted by 0.83 × 0.79 = 0.65 and that of P about Safe(R) is discounted by 0.786 × 0.75 = 0.59. However, this approach neglects the amount of evidence used to calculate trust in sources.

4.4 Evidence-based discounting

Within the evidence analysis domain, the information that we have to work with relates to past experiences with a specific agent (i.e., information source) gk where information received has proven reliable or unreliable according to some criteria (as would be captured in any trust assessment model). In other words, the amount of positive evidence we have for agent gk, namely rk, and the amount of negative evidence for that agent, namely sk. From this evidence, we calculate trustworthiness of gk, denoted as tk described in Section 2.2. When we receive opinion w_k^i from g_k, we discount it by t_k and add the resulting opinion w_k to our knowledge base. However, as explained before, additional discounting by factor x is required when w_k is in conflict with another opinion in the knowledge base. Discounting w_k by x implies discounting the original opinion w_k^i by t_k.x, which corresponds to revising the trustworthiness of w_k^i as t_k.x, by speculating about the trustworthiness of g_k regarding this single opinion. That is, even though the trustworthiness of g_k is t_k based on the existing evidence (t_k, s_k), it becomes t_k.x, for this specific opinion w_k^i; so, t_k.x effectively becomes the trust in w_k^i.
Here, we create a metric to measure how much we speculate about the trustworthiness of $\varrho_k$ regarding $w_k$.  

First, to decrease trust from $t_k$ to $t_k x_i$, we need additional negative evidence, which is called speculative evidence and denoted as $\rho_i$. Our intuition is that it is less likely for a trustworthy agent to present additional negative speculative evidence than it is for an untrustworthy agent, and thus the receipt of such evidence should be tempered by $(t_k)^{\kappa}$. Here, $t_k$ represents the distrust we have in agent $\varrho_k$; i.e. the likelihood that we will receive additional negative evidence given our experiences with the source. The calibration constant $\kappa \geq 0$ enables us to vary the influence that prior experience has on our prediction that an individual will present negative evidence in the future. If $\kappa = 0$, for example, we assume that all sources are equally likely to provide negative evidence. Now, using the Beta distribution formula for trust, we obtain:

$$t_k x_i = \frac{r_k + 1}{r_k + x_k + 2} = \frac{r_k + 1}{s_k + r_k + 2 + \rho_i \cdot (t_k)^{\kappa}}$$

Rearranging this for $\rho_i$ yields:

$$\rho_i = \frac{\nu_i (1 - x_i)}{x_i}$$

where $\nu_i = \frac{(r_k + s_k + 2) \kappa + 1}{(s_k + 1) \kappa}$.

To illustrate this, let us return to case II above in which agent C reports that road R is bombed by $G_1$. Using Equation 3, we can compute the total amount of speculative evidence necessary to discount $w_3$ and $w_4$ by $x_1$ and $x_2$, respectively. If we assume that $c_{12} x_1 + c_{21} x_2 = 1$, we have $x_2 = \frac{1 - c_{12} x_1}{c_{21}}$. Then, the total amount of speculative evidence (i.e. $\rho_1 + \rho_2$) can be formulated as a function of single variable $x_1$ by Equation 4, which is plotted in Figure 1. This function has a minimum at $x_1 = 0.5892$ in the interval $[0, 1]$, and the corresponding $x_2$ is 0.9607. That is, for a consistent knowledge base, trust in C’s opinion about roadBombedBy(R, $g_1$) should be reduced to 0.489 from 0.83, but the trust in the opinion of P about Safe(R) is reduced only slightly to 0.755 from 0.786. This reflects the relative level of positive and negative evidence we have from prior experience from both parties, and results in a consistent knowledge-based from which we can draw conclusions.

$$f(x_1) = \nu_1 (1 - x_1) + \nu_2 \left(1 - \frac{t_k (1 - c_{12} x_1)}{1 - c_{21} x_1}\right)$$

### Table 6: Extended ABox for case III

<table>
<thead>
<tr>
<th>Predicate</th>
<th>Asserted instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>roadBombedBy(R, $g_1$)</td>
<td>(0.67, 0.0, 0.247)</td>
</tr>
<tr>
<td>roadBombedBy$^-$ (R,$g_1$)</td>
<td>(0.67, 0.0, 0.247)</td>
</tr>
<tr>
<td>roadBombedBy(R)</td>
<td>(0.67, 0.0, 0.33)</td>
</tr>
<tr>
<td>roadBombedBy$^-$ (G_1)</td>
<td>(0.67, 0.0, 0.33)</td>
</tr>
<tr>
<td>Safe(R)</td>
<td>(0.63, 0.066, 0.304)</td>
</tr>
<tr>
<td>BombedRoad(R)</td>
<td>(0.2, 0.6, 0.2)</td>
</tr>
<tr>
<td>SabotagedRoad(R)</td>
<td>(0.801, 0.199)</td>
</tr>
</tbody>
</table>

### Table 7: After extra discounting for case III ($\kappa = 1$)

<table>
<thead>
<tr>
<th>Agent/ABox</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Safe(R)</td>
<td>(0.63, 0.066)</td>
</tr>
<tr>
<td>BombedRoad(R)</td>
<td>(0.2, 0.6, 0.2)</td>
</tr>
<tr>
<td>roadBombedBy(R, $g_1$)</td>
<td>(0.0342, 0.0034)</td>
</tr>
<tr>
<td>roadBombedBy$^-$ (G_1, R)</td>
<td>(0.0342, 0.0034)</td>
</tr>
<tr>
<td>roadBombedBy(R, $g_1$)</td>
<td>(0.0342, 0.0034)</td>
</tr>
<tr>
<td>roadBombedBy$^-$ (G_1)</td>
<td>(0.0342, 0)</td>
</tr>
<tr>
<td>BombedRoad(R)</td>
<td>(0.0342, 0.63)</td>
</tr>
<tr>
<td>roadBombedBy$^-$ (G_1)</td>
<td>(0.0342, 0.63)</td>
</tr>
<tr>
<td>roadBombedBy(R)</td>
<td>(0.0342, 0.63)</td>
</tr>
<tr>
<td>roadBombedBy$^-$ (G_1)</td>
<td>(0.0342, 0.63)</td>
</tr>
<tr>
<td>roadBombedBy$^-$ (G_1)</td>
<td>(0.0342, 0.63)</td>
</tr>
</tbody>
</table>

Until now, we considered only one conflict between two opinions. When we have multiple conflicts, they may interact in such a way that resolving one may also affect the resolution of another. To illustrate this, consider two new intelligence reports (case III):  

- A reports a bomb explosion on R with opinion (0.2, 0.6, 0.2)  
- $M_2$ informs $H_1$ that R is sabotaged with opinion (0.9, 0, 0.1).

The resulting ABox is shown in Table 6 and implies three relevant conflicts: $0.67 + 0.63 > 1, 0.67 + 0.1 > 1$, and $0.63 + 0.801 > 1$. Let us refer to (0.2, 0.6, 0.2) and (0.801, 0.199) as $w_3$ and $w_4$, respectively. We refer to the conflicting portions as $c_{31} = 0.6$ and $c_{42} = 0.801$. We also use $x_3$ and $x_4$ to refer to the additional discounting necessary for $w_3$ and $w_4$, respectively, to resolve the conflicts. The overall amount of speculative evidence necessary to resolve all of these relevant conflicts is computed as in Equation 5.

$$f(x_1) = \nu_1 (1 - x_1) + \nu_2 (1 - x_2) + \nu_3 (1 - x_3) + \nu_4 (1 - x_4)$$

such that

- $0 \leq c_{12} x_1 + c_{21} x_2 \leq 1$ and
- $0 \leq c_{13} x_3 + c_{31} x_1 \leq 1$
- $0 \leq c_{24} x_2 + c_{42} x_4 \leq 1$

Since these conflicts are relevant, we can write $x_2$, $x_3$, and $x_4$ in terms of $x_1$ if we set $c_{12} x_1 + c_{21} x_2 = 1$, $c_{13} x_1 + c_{31} x_3 = 1$, and $c_{24} x_2 + c_{42} x_4 = 1$. The resulting function is shown in Figure 2 and has a minimum at $x_1 = 0.0514$ in the interval [0, 1]. The other discounting factors are computed as $x_2 = x_3 = 1$ and $x_4 = 0.043$ in the same interval. That is, trust in the opinion of C about roadBombedBy(R, $g_1$) is reduced to 0.0427 and trust in the opinion of P about SabotagedRoad(R) is reduced to 0.0338. The ABox and the computed interpretations after extra discounting is shown in Table 7.

We generalise this approach for any number of conflicts with arbitrary relations. Assume we have a set of conflicting opinions \{ \langle w_1, w_2 \rangle, \ldots, \langle w_m, w_n \rangle \} and, derived from trust evidence about agents, coefficients \{ \nu_1, \nu_2, \ldots, \nu_m, \nu_n \}. To determine the optimum discounting factors \{ x_1, x_2, \ldots, x_m, x_n \} for these opinions, we construct the following optimisation problem with a multivariate non-linear objective function and linear constraints.

$$\arg \min_{\mathbf{x}} f(\mathbf{x})$$

where

$$f(\mathbf{x}) = \nu_1 (1 - x_1) + \nu_2 (1 - x_2) + \cdots$$

$$\nu_m (1 - x_m) + \nu_n (1 - x_n)$$

such that

- $0 \leq x_1 \leq 1.0 \leq x_j \leq 1.0 \ldots$ and
- $0 \leq c_{ij} x_i + c_{ji} x_j \leq 1.0 \ldots$

Existing constrained non-linear programming methods can be used to solve this problem in order to estimate the best discounting factors. There are various techniques that may be used including Interior-Point and Active-Set algorithms. In this work, we use Interior-Point approximation. Details of these methods are out of the scope of this paper and can be found elsewhere [14].

In this section we have formalised the problem of computing additional discounting factors for opinions received about the world from other agent so that we may formulate a consistent SDL-Lite
knowledge-base from which we can draw reliable conclusions. We have presented a number of approaches to the resolutions of inconsistencies between opinions including an optimisation-based approach, evidence-based discounting. Next, we evaluate these approaches with respect to their robustness in the face of liars.

5. EVALUATION

We have evaluated our approach through a set of simulations. In each simulation, we define the domain by randomly generating a $\mathcal{DL-Lite}$ TBox that contains 100 concepts and roles, as well as axioms over those, e.g., $B_1 \subseteq B_2$ and $B_2 \subseteq \neg \exists R_3$. For each role or concept, there is one information source that provides opinions about its instances, e.g., $B_1(a):(0.8, 0.2)$ and $R_3(a, b):(0.5, 0.1, 0.4)$. There are 10 information sources in total, each is an expert on 10 concepts and roles, and provides its opinions about those.

In our simulations, we assume there is one information consumer that uses the information from sources to make decisions. Each simulation is composed of 10 iterations. At each iteration $t$, the consumer needs to gather information about an individual $a$. We generate ground truth about $a$, which is composed of one assertion about $a$ for each concept and role with an associated opinion. Each information source knows the ground truth only about the concepts and roles of their expertise. However, they may not provide the ground truth to the consumer when it is requested. Behaviours of the information sources are determined by their behavioural type, which are summarised as follows:

- Honest: Most of the time, this type of sources provide the ground truth about the assertion of their expertise with small Gaussian noise $N(0, 0.01)$. With probability $P_h$, honest sources behave like malicious ones and provide bogus information.

- Malicious: This type of sources aim at misleading the information consumer by providing bogus opinions. More specifically, given $(b, d, u, s)$ is the ground truth about an assertion, a malicious source provides the opinion $\langle \text{abs}(\epsilon_1), 0.9 + \epsilon_2, u, s \rangle$ if $b \approx d$; otherwise it provides the opinion $\langle d + \epsilon_1, b + \epsilon_2, u, s \rangle$, where $\epsilon_1, \epsilon_2 \in [-0.05, 0.05]$. There are two types of malicious sources, which are defined as follows:
  
  i. Simple liars: they always provide bogus opinions.
  ii. Strategic liars: they behave like honest sources to build trust and then provides bogus information exploiting the built trust. After providing misleading information to the consumer, they change their identity to avoid negative evidence against them.

After collecting opinions about different assertions from information sources, the information consumer uses its trust in these sources to discount these opinions and uses the proposed reasoning mechanisms for $\mathcal{DL-Lite}$ to compute interpretations. Ideally, these interpretations should be close to the ground truth if all sources are accurate and their trustworthiness is modelled correctly. If there are some malicious sources, there may be conflicts in the collected information. In the case of conflicts, the consumer resolve the conflicts using Naive Deleting (NDL), Trust-based Deleting (TDL), Trust-based Discounting (TDC), or Evidence-based Discounting (EDC) with $\kappa = 1$. In NDL, all conflicting opinions are deleted from the knowledge base to resolve the conflicts. The consumer computes the interpretations for concept and role assertions related to $a$, after resolving the conflicts if any. Then, we measure the performance as the mean absolute error in the computed interpretations. Let $(b, d, u, s)$ be the ground truth and $(b', d', u', s')$ be the computed interpretation for assertion $B(a)$, then the absolute error in the interpretation is computed as $\text{error}(B(a)) = \text{abs}(b - b') + \text{abs}(d - d')$, where $\delta_b = b' - b$ and $\delta_d = d' - d$. For instance, if the ground

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Simple liars ($R_{\text{liar}} = 0.5$ and $P_b = 0$)}
\end{figure}

truth about $B(a)$ is $(0.9, 0.05, 0.05)$, but the computed interpretation is $(0.05, 0.9, 0.05)$, then the error would be 1.7.

At the end of each iteration, the consumer learns the ground truth and updates the trustworthiness of the information sources with new evidence $(r^i, s^i)$ computed as in Equation 7, which is based on the intuition that the information is still useful if it has a small amount of noise or is slightly discounted.

\begin{equation}
(r^i, s^i) = \begin{cases}
(0, 1), & \text{if } \delta_b > 0.1 \text{ or } \delta_d > 0.1 \\
(1, 0), & \text{if } -0.1 \leq \delta_b \leq 0.01 \text{ and } -0.1 \leq \delta_d \leq 0.01 \\
(0, 0), & \text{otherwise},
\end{cases}
\end{equation}

Each of our simulations are repeated 10 times and our results are significant based on t-test with a confidence interval of 0.95.

Without any evidence, the trustworthiness of sources is computed as 0.5. Thus, there are no conflicts in the beginning of our simulations. If all sources have deterministic behaviours, i.e., malicious sources are simple liars and $P_b = 0$, then trustworthiness of sources are easily modelled over time and the opinions from liars are significantly discounted. In such settings, conflicts are totally avoided and information consumers using either of the four proposed methods have the same level of success. Figure 3 shows an example of this setting where honest sources always provides the truth ($P_b = 0$) and malicious sources are simple liars. Here, the ratio of liars ($R_{\text{liar}}$) is 0.5, i.e., half of the sources are malicious.

When honest sources provide bogus information occasionally, the conflicts may arise in the knowledge base of the consumer, because the information from these sources are not significantly discounted. Figure 4 shows our results for $R_{\text{liar}} = 0.5$ and $P_b = 0.1$, where all malicious sources are simple liars. In this setting, NDL leads to significant errors in the computed interpretations. While TDL does much better than NDL, it is outperformed by discounting based approaches TDC and EDC. Both of these approaches have similarly good performance though TDC does slightly better.

Simple liars may not be enough to model malicious sources in real life. That is why we change the type of malicious sources

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{Simple liars ($R_{\text{liar}} = 0.5$ and $P_b = 0.1$)}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure5.png}
\caption{Strategic liars ($R_{\text{liar}} = 0.5$ and $P_b = 0.1$)}
\end{figure}

\begin{table}[h]
\centering
\begin{tabular}{|c|c|}
\hline
Method & Mean Absolute Error \\
\hline
Naive Deleting & 0.63 \\
Trust-based Deleting & 0.58 \\
Trust-based Discounting & 0.56 \\
Evidence-based Discounting & 0.54 \\
\hline
\end{tabular}
\caption{Comparison of different methods}
\end{table}
fully handle highly misleading information in challenging settings. Through simulations, we show that our approach can successfully identify and resolve conflict among information providers on the web. In Proceedings of The 9th International Conference on Knowledge and Data Discovery, 2007.

7. REFERENCES


