Assessing Trust in Uncertain Information using Bayesian Description Logics

Achille Fokoue
IBM Watson Research Center
Yorktown Heights, NY, USA
achille@us.ibm.com

Mudhakar Srivatsa
IBM Watson Research Center
Yorktown Heights, NY, USA
msrivats@us.ibm.com

Rob Young
Dstl
Wiltshire, UK
riyoung@dstl.gov.uk

ABSTRACT

Decision makers (humans or software agents alike) are faced with the challenge of examining large volumes of information originating from heterogeneous sources with the goal of ascertaining trust in various pieces of information. In this paper we argue (using examples) that traditional trust models are limited in their data model by assuming a pair-wise numeric rating between two entities (e.g., eBay recommendations, Netflix movie rating, etc). We present a novel trust computational model for rich, complex and uncertain information encoded using Bayesian Description Logics. We present security and scalability tradeoffs that arise in the new model, and the results of an evaluation of the first prototype implementation under a variety attack scenarios.

Categories and Subject Descriptors: C.2.0 [General]: Security and protection

General Terms: Trust Assessment

Keywords: Robust Trust Models, Uncertain Information, Shilling, Bad-mouthing, Scalable Trust Assessment

1. INTRODUCTION

Decision makers (humans or software agents alike) relying on information available on the web are increasingly faced with the challenge of examining large volumes of information originating from heterogeneous sources with the goal of ascertaining trust in various pieces of information. Several authors have explored various trust computation models (e.g., eBay recommendation system [9], NetFlix movie ratings [8], EigenTrust [6], PeerTrust [10], etc.) to assess trust in various entities. A common data model subsumed by several trust computation models (as succinctly captured in Kuter and Golbeck [7]) is the ability of an entity to assign a *numeric* trust score to another entity (e.g., eBay recommendation, Netflix movie ratings, etc.). Such pair-wise numeric ratings contribute to a (dis)similarity score (e.g., based on $\mathcal{L}_1$ norm, $\mathcal{L}_2$ norm, cosine distance, etc.) which is used to compute personalized trust scores (as in PeerTrust) or recursively propagated throughout the network to compute global trust scores (as in EigenTrust).

A pair-wise numeric score based data model may impose severe limitations in several real-world applications. For example, let us suppose that information sources $\{S_1, S_2, S_3\}$ assert axioms ϕ_1 = *all men are mortal*, ϕ_2 = *Socrates is a man* and ϕ_3 = *Socrates is not mortal* respectively. While there is an obvious conflict when all the three axioms are put together, we note that: (i) there is no pair-wise conflict, and (ii) there is no obvious numeric measure that captures (dis)similarity between two information sources.

This problem becomes even more challenging because of uncertainty associated with real-world data and applications. Uncertainty manifests itself in several diverse forms: from measurement errors (e.g., sensor readings) and stochasticity in physical processes (e.g., weather conditions) to reliability/trustworthiness of data sources; regardless of its nature, it is common to adopt a probabilistic measure for uncertainty. Reusing the *Socrates* example above, each information source S_i may assert the axiom ϕ_i with a certain probability $p_i = 0.6$. Further, probabilities associated with various axioms need not be (statistically) independent.

In such situations, the key challenge is to develop trust computation models for rich (beyond pair-wise numeric ratings) and uncertain (probabilistic) information. In doing so, the trust computation model must address two issues:
Security: The trust computation model must be robust to *shilling* (malicious entities promoting each other's trust scores) and *bad-mouthing* (malicious entities attempting to decrease the trust scores for honest entities). In this paper we consider various attack scenarios including malicious collusions, imperfect honest entities, information asymmetry amongst sources (sparsity of information), and attempts to milk the trust model (e.g., oscillatory behavior).
Scalability: The trust computation model must scale with the volume of information and the number of information sources. In this paper we highlight security-scalability tradeoffs that arise in the new trust computation model. We demonstrate scalability by evaluating our prototype implementation on a publicly available UOBM dataset (74,000 statements).

2. TRUST COMPUTATION MODEL

2.1 Data Model

While prior trust models allow sources to make statements on pair-wise trust scores (e.g., entity i assigns a numeric score $s \in [0, 1]$ to entity j), we allow sources to make richer and more expressive statements encoded in Bayesian Description Logics [2] (BDL)[1] with axioms of the form $\phi : X$ where ϕ is a classical axiom (in Description Logics (DL [1])) that is annotated with a boolean random variable from a

[1]BDL is a simple probabilistic extension of Description Logics, the foundation of Semantic Web.

CCS'10, October 4–8, 2010, Chicago, Illinois, USA.
ACM 978-1-4503-0244-9/10/10.

Bayesian network [4]. Intuitively, $\phi : X$ can be read as follows: the axiom ϕ holds when the Boolean random variable X is true[2]. Dependencies between axioms (e.g., $\phi_1 : X_1$ and $\phi_2 : X_2$) are captured using the Bayesian network that represents a set of random variables (corresponding to the annotations; e.g., X_1, X_2) and their conditional probability distribution functions (e.g., $Pr(X_2|X_1)$).

The following example illustrates how this formalism can be used to describe road conditions influenced by probabilistic events such as weather conditions:
$SlipperyRoad \sqcap OpenedRoad \sqsubseteq HazardousCondition$ (an open slippery road is hazardous), $Road \sqsubseteq SlipperyRoad : Rain = true$ (a road is slippery when it is raining), $= \{Road(route9A)$ ($route9A$ is a road), $OpenedRoad(route9A) : TrustSource = true$ ($route9A$ is open when then it is reported by a trusted source). In this example, the Bayesian network BN consists of three self-explanatory variables: $Rain$, $TrustSource$, and $Source$. The probabilities specified by BN are as follows: $Pr_{BN}(TrustSource = true|Source =$ '$Mary'$) = 0.8, $Pr_{BN}(Rain = true) = 0.7$,$Pr_{BN}(Source =$ '$John'$) = 1, $Pr_{BN}(TrustSource = true|Source =$ '$John'$) = 0.5.

Informally, probability values computed through the Bayesian network 'propagate' to the 'DL side' as follows. Each assignment v of all random variables in BN (e.g., $v = \{Rain =$ true, $TrustSource =$ false, $Source =$ 'John'$\}$) corresponds to a primitive event ev (or a scenario). A primitive event ev is associated, through BN, to a probability value p_{ev} and a classical DL KB K_{ev}[3] which consists of all classical axioms annotated with a compatible probabilistic annotation (e.g., $SlipperyRoad \sqcap OpenedRoad \sqsubseteq HazardousCondition$, $Road \sqsubseteq SlipperyRoad, Road(route9A)$). The probability value associated with the statement ϕ (e.g., $\phi = HazardousCondition(route9A)$) is obtained by summing p_{ev} for all ev such that the classical KB K_{ev} entails ϕ (e.g., $Pr(HazardousCondition(route9A)) = 0.35$).

2.2 Trust Assessment

Our approach offers a trust computation model over uncertain information (encoded as BDL axioms). Intuitively, our approach allows us to compute a degree of inconsistency over a probabilistic knowledge base. We note that inconsistencies correspond to conflicts in information items reported by one or more information sources. Our approach assigns numeric weights to the degree of inconsistency using the *possible world* semantics.

Revisiting the *Socrates* example, three probabilistic axioms $\phi_i : p_i$[4] correspond to eight possible worlds (the power set of the set of axioms without annotations) corresponding to $\{\{\phi_1, \phi_2, \phi_3\}, \{\phi_1, \phi_2\}, \cdots, \emptyset\}$. For instance, the possible world $\{\phi_1, \phi_2\}$ corresponds to a world wherein *all men are mortal*, and *Socrates is a man*. Each possible world has probability measure that can be derived as joint probability distributions over the random variables in the Bayesian network. For instance, the probability of a possible world $\{\phi_1, \phi_2\}$ is given by $p_1 * p_2 * (1 - p_3)$.

In the presence of inconsistencies, our approach extracts justifications − minimal sets of axioms that together imply an inconsistency [5]. Our trust computation model propagates the degree of inconsistency as blames (or penalties) to the axioms contributing to the inconsistency via justifications as follows. First, we compute a probability measure for each justification as the sum of the probabilities associated with possible worlds in which the justification holds (namely, all the axioms in the justification are present). Second, we partition the degree of inconsistency across all justifications; for instance, if a justification J_1 holds in 80% of the possible worlds then it is assigned four times the blame as a justification J_2 that holds in 20% of the possible worlds. Third, we partition the penalty associated with a justification across all axioms in the justification.

2.2.1 Security-Scalability Tradeoff

A naive implementation of our trust computation model requires *all* justifications. While computing a justification is an easy problem, exhaustively enumerating all possible justifications is known to be hard problem [5]. For scalability reasons, the trust computation model uses K justifications; however, such justifications must be sampled at random. An unbiased sample of justifications ensures that the malicious entities cannot game the trust computation model; say, selectively hide justifications that include axioms from malicious entities (and thus evade penalties) from the sampling process.

We formulate exhaustive enumeration of justifications as Reiter's Hitting Set Tree (HST) traversal problem [5] (see Figure 1). One can find the first K conflicts by exploring the Reiter's Hitting Set Tree (HST) until K distinct justifications are found. The problem with this approach is that nodes in the HST are not equally likely to be selected with such a scheme: the probability $\pi(v_d)$ of a node v_d in a path $< v_0 v_1 ... v_d >$ to be selected is $\pi(v_d) = \prod_{0 \leq i < d}(1/|v_i|)$, where $|v_i|$ denotes the number of axioms in the justification v_i. Since the bias can be precisely quantified, one can obtain an unbiased sample as follows. We select K nodes in the HST by traversing the HST in any order, but each time a node v_i is encountered, it is selected with probability $\min(\frac{\beta}{\pi(v_i)}, 1)$, where β is a strictly positive real number.

Our trust computation model tradeoffs security with scalability via β: a large β improves efficiency at the cost of introducing more bias (more vulnerable to attacks); while small β increases the cost of traversing the HST while lowering the bias[5] (less vulnerable to attacks).

3. EXPERIMENTAL EVALUATION

We have developed a prototype of our trust assessment system by implementing a probabilistic extension, PSHER, to our publicly available highly scalable DL reasoner SHER [3]. We empirically evaluated the efficacy of our scheme (on a publicly available UOBM dataset) when malicious sources use an oscillating behavior to milk the trust computation model and when honest sources are faced with measurement errors (high uncertainty) or commit honest mistakes.

In our experiments, we considered 4 types of information sources:
Perfect honest sources (PHS) whose axioms are taken from the UOBM KB before the introduction of inconsistencies.
Purely malicious sources (PuMS) whose axioms are selected

[2]If X is *null* then ϕ always holds
[3]K_{ev} was informally referred to as a 'possible world'
[4]$\phi_i : p_i$ is a shorthand notation for $\phi_i : X_i$ and $Pr(X_i = true) = p_i$ for some independent random variable X_i
[5]One can obtain a perfectly unbiased sample by setting, $0 < \beta < \min_{v_i}\{\pi(v_i)\}$

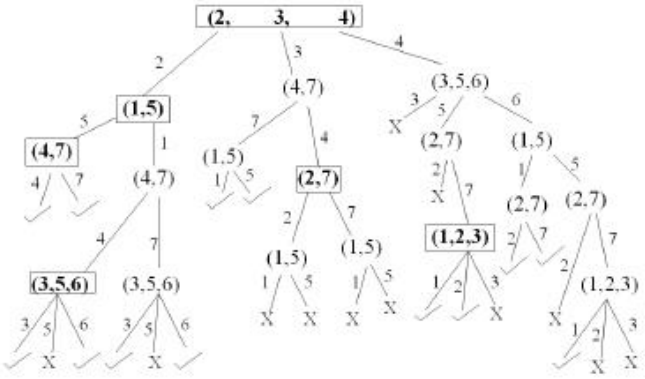

Figure 1: Computing all justifications using Reiter's Hitting Set Tree Algorithm from [5]

from the ones added to UOBM KB in order to create inconsistencies.

Imperfect honest sources (IHS) have the majority of their axioms (more than 90%) from the UOBM KB before the introduction of inconsistencies. They allow us to simulate the behavior of our approach when honest sources are faced with measurement errors or commit honest mistakes.

Partially malicious sources (PaMS) are such that between 10% to 90% of their axioms are selected from the axioms added to UOBM KB to create inconsistency. They are primarily used to simulate the behavior of our approach when malicious sources use an oscillating behavior to milk our trust computation scheme.

Axioms were randomly assigned to various sources without violating the proportion of conflicting vs. non-conflicting axioms for each type of source. Note that not all axioms in the original UOBM KB end up being part of an inconsistency, which introduces an asymmetry in information source's knowledge (e.g., a malicious source is not assumed to have complete knowledge of all axioms).

Our experiment simulates an oscillating scenario where all four types of sources are present: 30% PHS, 20% PuMS, 30% IHS and 20%PaMS. The malicious sources alternate periods where they assert incorrect axioms, contradicting axioms asserted in the same period by other sources, with periods in which they assert only correct axioms. Figure 2 shows how our scheme correctly separates the 4 types of sources as expected. In absence of prior knowledge, the trust values of partially malicious sources (PaMS) and honest sources drop significantly at the first period in which incorrect axioms are stated. However, malicious sources, which due to information asymmetry, can only contradict limited set of statements from honest sources, never recover significantly, while honest sources quickly improve their trust values by asserting more axioms not involved in conflicts. The negative impact on honest sources can be reduced considerably through axiom duplication and trust priors.

4. SUMMARY

In this paper[6], we have introduced a new trust computation model for rich, complex and uncertain information by leveraging the expressiveness of Bayesian Description Logics. We have demonstrated the robustness of the proposed framework under a variety of scenarios, and shown how duplication of assertions across different sources as well as prior knowledge of the trustworthiness of sources can further enhance it.

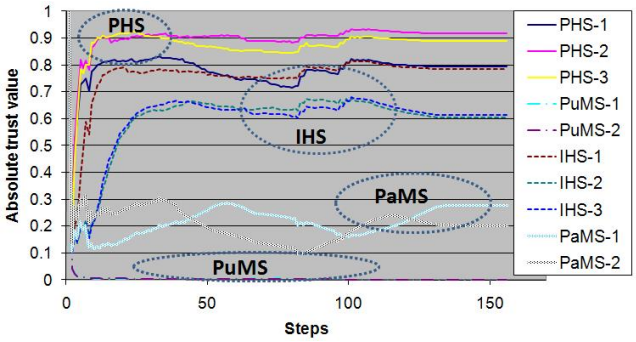

Figure 2: Oscillating experiment - 30% PHS, 20% PuMS, 30% IHS & 20% PaMS

5. REFERENCES

[1] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider. *The Description Logic Handbook*. Cambridge University Press, 2003.

[2] C. D'Amato, N. Fanizzi, and T. Lukasiewicz. Tractable reasoning with bayesian description logics. In *Scalable Uncertainty Management (SUM08)*, pages 146–159, 2008.

[3] J. Dolby, A. Fokoue, A. Kalyanpur, E. Schonberg, and K. Srinivas. Scalable highly expressive reasoner (sher). *J. Web Sem.*, 7(4):357–361, 2009.

[4] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference and Prediction. In *Springer Series in Statistics*, 2009.

[5] A. Kalyanpur. *Debugging and Repair of OWL-DL Ontologies*. PhD thesis, University of Maryland, 2006.

[6] S. Kamvar, M. Schlosser, and H. Garcia-Molina. EigenTrust: Reputation management in P2P networks. In *WWW*, 2003.

[7] U. Kuter and J. Golbeck. SUNNY: A New Algorithm for Trust Inference in Social Networks, using Probabilistic Confidence Models. In *AAAI-07*, 2007.

[8] Netflix. Netflix Prize. http://www.netflixprize.com/.

[9] J. B. Schafer, J. Konstan, and J. Riedl. Recommender Systems in E-Commerce. In *ACM Electronic Commerce*, 1999.

[10] L. Xiong and L. Liu. Supporting reputation based trust in peer-to-peer communities. In *IEEE TKDE, Special Issue on P2P Data Management, vol. 71, 16(7)*, July 2004.

[6]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.