Joslyn, Cliff A and Weaver, Jesse: (2013) "Focused Belief Measures for Uncertainty Quantification in High Performance Semantic Analysis", in: 2013 Conf. on Semantic Technologies in Defense and Security (STIDS 2013), CEUR-WS, v. 1097, pp. 18-24, http://ceur-ws.org/Vol-1097

Focused Belief Measures for Uncertainty Quantification in High Performance Semantic Analysis

Cliff Joslyn
National Security Directorate
Pacific Northwest National Laboratory
Seattle, WA 98109
Email: cliff.joslyn@pnnl.gov

Jesse Weaver
Fundamental and Computational Sciences Directorate
Pacific Northwest National Laboratory
Richland, WA 99354
Email: jesse.weaver@pnnl.gov

***Abstract*—In web-scale semantic data analytics there is a great need for methods which aggregate uncertainty claims, on the one hand respecting the information provided as accurately as possible, while on the other still being tractable. Traditional statistical methods are more robust, but only represent distributional, additive uncertainty. Generalized information theory methods, including fuzzy systems and Dempster-Shafer (DS) evidence theory, represent multiple forms of uncertainty, but are computationally and methodologically difficult. We require methods which provide an effective balance between the complete representation of the full complexity of uncertainty claims in their interaction, while satisfying the needs of both computational complexity and human cognition. Here we build on Jøsang's subjective logic to posit methods in focused belief measures (FBMs), where a full DS structure is focused to a single event. The resulting *ternary* logical structure is posited to be able to capture the *minimally sufficient* amount of generalized complexity needed at a *maximum* of computational efficiency. We demonstrate the efficacy of this approach in a web ingest experiment over the 2012 Billion Triple dataset from the Semantic Web Challenge.**

I. Introduction

Many analytic domains face the problem of determining the veracity of claims from multiple sources. The problem can be further complicated by the presence of large numbers of sources asserting large numbers of propositions over short periods of time. Examples include intelligence gathering and sensor networks. Such problems are only exacerbated on the web with the constituent heterogeneity of data, and then again especially when brought to web scale.

While there are various logics for dealing with inconsistency or uncertainty [3], [13], [14], to our knowledge, none have achieved significant uptake in computational systems for large data. Traditional statistical uncertainty representation (UR) models fail to represent complex uncertainty situations requiring imprecision or other forms of ambiguous judgements. So-called Generalized Information Theory (GIT) approaches [12] such as fuzzy and Dempster-Shafer (DS) can represent these complexities, but at the cost of high computational expense. Massive streaming or ingest problems in the semantic web require UR strategies which provide both representation of ambiguity and computational efficiency.

Here we present Focused Belief Measures (FBMs), an adaptation of Jøsang's subjective logic (SL) [10], as a candidate to play this role. By modifying DS to focus on specific events in a complex space, FBMs can model logical combinations of complex beliefs involving imprecision, ambiguity, or total ignorance using *linear* algorithms. This compromise promises to support the *minimal* amount of generalized complexity which may be nonetheless sufficient, but at a *maximum* of computational efficiency.

We begin by introducing FBMs in the context of both SL and DS. We then demonstrate its utility in a web analytics experiment involving the evaluation of a large RDF graph drawn from the 2012 Semantic Web Challenge.

II. Focused Belief Measures (FBMs)

UR methods and formalisms are legion, primarily rooted in probability theory, logic, or their combination (e.g. [5]). For decades, UR researchers have struggled with two competing imperatives. On the one hand, traditional UR methods, including probabilistic (statistical) and logical approaches require closed-world assumptions. For probability, we require knowledge about likelihood distributions over a set of entities which are both exhaustive and mutually exclusive in order to guarantee mathematical additivity: summation of all modeled probabilities to 1. Representing total uncertainty requires assuming a uniform distribution over these choices. Similarly, traditional logic represents only the two states for true (A) and false ($\sim A$, here taking $\sim$ for negation), according to an excluded middle axiom.

The large range of GIT methods, including fuzzy systems, DS evidence theory, random sets, imprecise probabilities [16], and many others, support open world situations by allowing some form of *third option* or *remainder* between True and False, or non-additive, imprecise "overlap" between non-disjoint options. They do this in different ways, but the basic concept is the same, to generalize traditional approaches by relaxing certain axioms, such as allowing probabilities to sum to more or less than 1, or to recognize truth values "between" True and False (different kinds of "Maybe"), in

order to represent more complex uncertainty structures other than probabilistic likelihoods. In this way one can represent inherent vagueness or imprecision of events, or second- or higher-order uncertainties about other uncertainties, reflecting the veracity of the information source, the confidence or likelihood of claims, the uncertainty about a claim, or other open world situations where "something else" we didn't think about could occur.

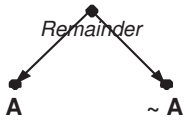

Fig. 1. In this simplest case, generalized information theories support imprecise choices by always including a "third option" or "remainder", allowing positive weight to "neither A nor $\sim A$.

These generalized mathematical structures are inherently hierarchical, since this "remainder" "stands above" its constituent choices. Fig. 1 shows the absolute simplest such case, where "neither A nor $\sim A$" is our "third choice" standing above A and $\sim A$ themselves. This remainder can be used to represent *total* uncertainty or imprecision when positive weight is given to the remainder, but no weight is given to either of the two specific choices.

Fig. 1 actually shows a tiny lattice structure, and the resulting methods require lattice-based computations arising from non-additivity. For example, classical probability has the condition $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$, which is a fully modular function on the subset lattice, allowing relatively simple calculations. But in non-additive formalisms, we have $\Pr(A \cup B) \gtrless \Pr(A) + \Pr(B) - \Pr(A \cap B)$, which is sub- or super-modularity [6]. Modularity allows probabilities of "bigger" events to be calculated from those of "smaller", so non-modularity forces a high computational price. In big data, semantic web environments with massive ingest and streaming input applications, we need methods for representing such hybrid uncertainty, but which are both expressive and tractable.

Our FBM approach is built on Jøsang's SL [10], which is in turn based on DS. Consider a decision problem, like whether Alice, Bob, or Carol committed a crime. In probability, we need $P = p(A) + p(B) + p(C) = 1$ to hold. If $P > 1$, then we have conflict and have to renormalize; while if $P < 1$, then we have a *remainder*, represented as an uncertainty $U = 1 - P > 0$ leftover. In DS theory, we represent general probabilistic uncertainty by giving probabilities m not just to each of these $n = 3$ disjoint events $A, B, C \in \Omega$, but to each of the 2^n *subsets* $R \subseteq \Omega$ of such events. Formally, we have

$$m\colon 2^\Omega \to [0,1], \quad m(\emptyset) = 0, \quad \sum_{R \subseteq \Omega} m(R) = 1.$$

We identify any $R \subseteq \Omega$ with $m(R) > 0$ as *focal*.

This supports modeling of imprecision and ignorance together with likelihood by assigning values to the completely imprecise event $m(\Omega)$, down to the most precise singletons $m(\{A\})$, and everything in between, including *composite* events like $m(\{A, B\})$ for "Alice and/or Bob did it".

The resultant **belief measures**

$$b\colon 2^\Omega \to [0,1], \quad b(R) = \sum_{S \subseteq R} b(S)$$

on any subset $R \subseteq \Omega$ capture a mixture of likelihood and imprecision, since claims m about subsets R cannot necessarily be disambiguated to knowledge about their constituent elements $\omega \in R$. But considering (the middle of) Fig. 2 compared to Fig. 1, we see that we now need to support the full Boolean lattice representing the power set 2^Ω of all subsets. This comes at a huge computational cost, since we now have to work with 2^n rather than n claims, and moreover their interaction within the lattice structure.

But Jøsang [10] has noted that if we *focus* attention to a *particular* composite event $R \subseteq \Omega$ (like $\{A, B\}$=Alice and/or Bob did it), we can reduce the complexity to just three *disjoint* groups of subsets:

1) Those (like $\{A\}$=Alice did it) completely within R supporting R itself;
2) Those (like $\{C\}$=Carol did it) completely disjoint from R supporting $\sim R$ (now taking $\sim R = \Omega \setminus R$ for set complement); and
3) The remainder (like $\{B, C\}$=Bob and/or Carol did it), providing information contradictory to or ambiguous with respect to *both* R and $\sim R$.

These three groups reduce to a single "opinion" vector

$$w(R) = \langle b(R), d(R), u(R) \rangle ,$$

where in addition to $b(R)$ as the **belief** of R, we have

$$d(R) = b(\sim R) = \sum_{S \subseteq \sim R} m(S)$$

as the belief of $\sim R$, that is the **disbelief** of R[1]. Since $b(R), d(R) \in [0,1]$ and $b(R) + d(R) \leq 1$, this allows us to define

$$u(R) = \sum_{S\colon \emptyset \neq S \cap R \neq S} m(S) = 1 - b(R) - d(R)$$

to serve elegantly as our generalized remainder, or **uncertainty** of R.

As so specified, $b(R) + d(R) + u(R) = 1$. Thus b, d, and u exhaust all the options concerning R and $\sim R$, but do so while *including* a representation of the "remainder", u, which is about "neither R nor $\sim R$". This reflects the fact that while,

[1]We depart from Jøsang in using this formulation for d, rather than his (equivalent) $d(R) = \sum_{S \cap R \neq \emptyset, S \not\subseteq R} m(S)$. His is both formally more complex and conceptually less cogent, since it does not capture the sense in which b and d represent the overall belief in the set R and its "opposite". Since our choice is to cast both b and d as beliefs, just in the set R and its opposite $\sim R$, and additionally since our formulation does not rely on anything inherently either "subjective" or "logical", we choose to identify our formulation as **focused belief measures** rather than Jøsang's "subjective logic".

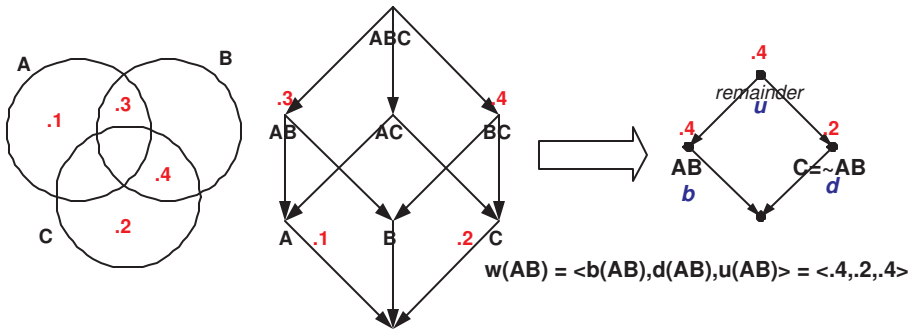

Fig. 2. Proabilities about Alice, Bob, and Carol, *and their combinations*, need $2^3 - 1 = 7$ assessments. When focused on $\{A, B\}$, we reduce to three.

for any $R \subseteq \Omega$, the two *sets* R and $\sim R$ partition Ω, it is rather the three *classes* (sets of subsets)

$$\{S \subseteq R\}, \quad \{S \subseteq \sim R\}, \quad 2^\Omega \setminus (\{S \subseteq R\} \cup \{S \subseteq \sim R\})$$

which partition the power set 2^Ω. It is this third class which is our remainder.

Consider opinions $w_A(R)$, $w_A(S)$, and $w_B(S)$ as opinions from information sources A and B about propositions R and S. Also let $w_A(B)$ be source A's opinion of source B. Jøsang then provides a series of algebraic operators for different combinations, including:

- **Conjunction**

$$w_A(R \wedge S) = \Big\langle b_A(R)b_A(S), \\ d_A(R) + d_A(R) - d_A(R)d_A(S), \\ b_A(R)u_A(R) + u_A(R)b_A(S) + \\ u_A(R)u_A(S) \Big\rangle$$

  and **disjunction**

$$w_A(R \vee S) = \Big\langle b_A(R) + b_A(R) - b_A(R)b_A(S), \\ d_A(R)d_A(S), \\ d_A(R)u_A(R) + u_A(R)d_A(S) + \\ u_A(R)u_A(S) \Big\rangle$$

  opinions about two different propositions by the same source;
- A parallel **consensus** operator

$$w_A(R) \oplus w_B(R) = \left\langle \frac{b_A u_B + b_B u_A}{u_A + u_B - u_A u_B}, \right. \\ \frac{d_A u_B + d_B u_A}{u_A + u_B - u_A u_B}, \\ \left. \frac{u_A u_B}{u_A + u_B - u_A u_B} \right\rangle$$

  expressing the opinion of one proposition R by two sources A, B; and
- A series **discounting** operator

$$w_A(B) \otimes w_B(S) = \Big\langle b_A(B)b_B(S), \\ b_A(B)d_B(S), \\ d_A(B) + u_A(B) + b_A(B)u_B(S) \Big\rangle$$

  expressing the discounted opinion about a base opinion $w_B(S)$ in light of another opinion $w_A(B)$, which we take to be A's opinion about the agent B *expressing* the opinion $w_B(S)$.

Note the tradeoff that FBMs make. We are not representing the full complexity of all $2^n - 1$ possible combinations required by DS; but for any R, or collection of Rs, we are able to directly model $R, \sim R$, *and their remainder*, and while in a *minimal* way, with a *maximal* amount of computational efficiency: the huge advantage of these operators is that they are *linear* in the components b, d, u of the opinion vectors w. Given that we care about only k such events, then in realistic cases we have reduced the size of our problem space to $2k \ll 2^n - 2$ (that is, to $O(k)$ from order $O(2^n)$). We have also vastly improved user comprehensibility, since conceptualizing operations on linear vectors is far less challenging than the structure of hypercubic Boolean lattices. Thus logical combinations of complex situations can be represented easily and cheaply, while still representing our "third option".

This is shown even more strongly in Fig. 3, now the case for $n = 4$ basic choices, shown in the 4-dimensional hypercube (Boolean 4-lattice), and displaying the 4 focal sets. If we wished to track all $2^n - 1 = 15$ possible choices, then so be it, but consider instead that we were only interested in the $k = 3$ choices $\{A\}, \{A, C\}$, and $\{A, B, C\}$. Then we would only need to track the $3k = 9$ pieces of information in the opinion vectors

$$w(A) = \langle .1, .4, .5 \rangle, w(AC) = \langle .3, .4, .3 \rangle$$

$$w(ABC) = \langle .6, 0, .4 \rangle$$

(note that here we simplify notation so that e.g. $ABC = \{A, B, C\}$). As shown, we've replaced the need to store and compute on a 4-dimensional hypercube in exchange for three 2-dimensional hypercubes. FBMs are thus *ternary*, avoiding

limited binary reasoning with a third category to represent "complete ignorance", but also avoiding the full 2^n complexity of the set-based DS.

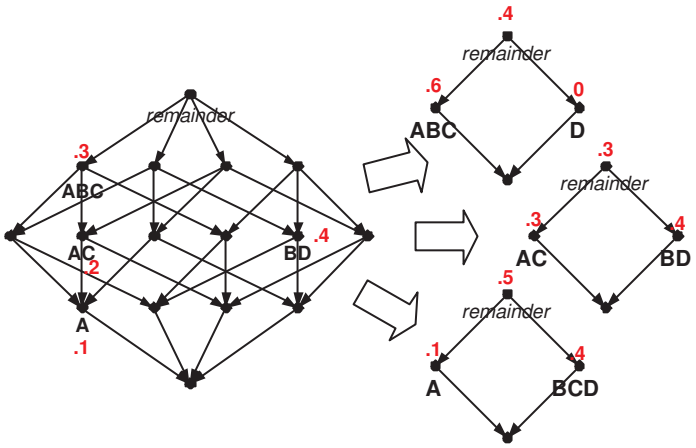

Fig. 3. (Left) Example focal structure for $n = 4$. (Right) Opinion structures for the three target sets $R = A, AC, ABC$.

III. Data Ingest Experiment

We next seek to demonstrate the value, feasability, and tractability of using FBMs in a data ingest experiment on the semantic web. The experiment reported on below is provisional, an initial foray into the basic operation of the FBM approach, but as specifically applied to web-based analytics. In particular, as will be described in more detail below, we use opinion vectors which are relatively constrained examples, being unequivocal for basic claims, but always with residual uncertainty in their aggregation. Further investigation can open the approach up to range over a wider array of values, seeking performance and sensitivity analyses.

A. FBM Problem Setup

Consider the following decision problem. Three data sources make claims about certain facts. Alice and Cindy assert p, but Bob says $\sim p$. The judge must determine whether he/she believes p is true. The judge generally believes Alice and Bob (Bob more than Alice) and thinks that Cindy is lying.

An FBM setup for this problem is shown in Fig. 4. First, the base claims made by Alice, Bob, and Cindy are unequivocally either true or fale (Jøsang calls these "dogmatic"), and thus lack the uncertainty parameter u. We have

$$w_A(p) = \langle 1, 0, 0 \rangle, \quad w_B(p) = \langle 0, 1, 0 \rangle, \quad w_C(p) = \langle 1, 0, 0 \rangle.$$

Next, the judge is inclined to believe Alice at about 80%, but it's not that her remaining 20% *disbelieves* Alice, but she is rather *uncertain* about Alice, not having further grounds to either believe of disbelieve her. We thus have $w(A) = \langle .8, 0, .2 \rangle$. The judge finds Bob convincing at 95%, even more than Alice, so $w(B) = \langle .95, 0, .05 \rangle$. Finally the judge is quite sure that Cindy is lying, but there is always the residual possibility otherwise, so $w(C) = \langle 0, .99, .01 \rangle$.

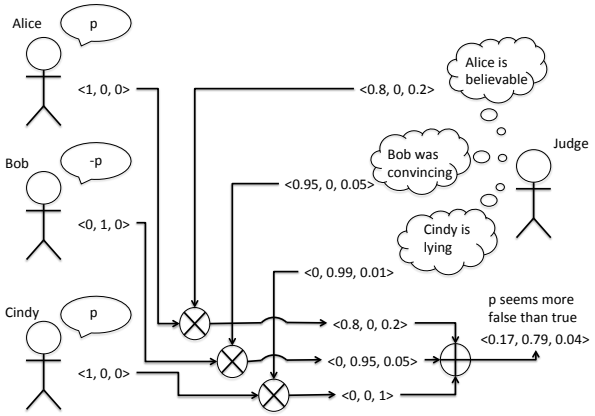

Fig. 4. An FBM problem setup: how to aggregate streaming opinions under source uncertainty?

Assume that Alice, Bob, and Cindy actually testify in order, modeling a streaming ingest operation. Initially, we have absolutely no knowledge about p, and thus the only valid choice is the totally uncertain opinion $w(p) = \langle 0, 0, 1 \rangle$. As an opinion $w_X(p)$ of p by source X arrives, we then update our opinion as:

$$w'(p) = w(p) \oplus (w(X) \otimes w_X(p)),$$

first discounting X's opinion of p with our opinion of X, and then aggregating with our prior opinion. We are thus building the consensus opinions as more data arrives from more sources, and the only state that needs to be saved is a single opinion for p.

After Alice, the judge's opinion of p is

$$\langle 0, 0, 1 \rangle \oplus (\langle 1, 0, 0 \rangle \otimes \langle 0.8, 0, 0.2 \rangle) = \langle 0.8, 0, 0.2 \rangle.$$

Then Bob testifies that p is absolutely false, or alternatively, that $\neg p$ is absolutely true. Now the judge's opinion of p is

$$\langle 0.8, 0, 0.2 \rangle \oplus (\langle 0, 1, 0 \rangle \otimes \langle 0.95, 0, 0.05 \rangle) = \langle 0.17, 0.79, 0.04 \rangle.$$

Bob has effectively changed the judge's mind. Finally, Cindy testifies that p is absolutely true, but the judge is nearly certain that she is lying. In the end, the judge's opinion of p is

$$\langle .17, .79, .04 \rangle \oplus (\langle 1, 0, 0 \rangle \otimes \langle 0, .99, .01 \rangle) = \langle .17, .79, .04 \rangle.$$

The judge's mind did not change at all as a result of Cindy's testimony. In the end, the judge is inclined to believe that p is false.

B. Billion Triple Challenge

Over the last 14 years, the Resource Description Framework (RDF) [2] has become a popular knowledge representation with mature research and tools to support it. It also has associated ontology languages such as RDF Schema (RDFS) [8] and the Web Ontology Language (OWL) [7] which allow

Triples	Number
Any FOAF-related triple	164.3M
With predicate `foaf:name`	7.8M
With predicate `foaf:knows`	17.3M

TABLE I
STATISTICS ABOUT FOAF TRIPLES IN BTC.

for declarative semantics that support reasoning tasks like inference and consistency checking. There are many efforts to expose data as RDF (e.g., Facebook [17], Data.gov [4], biomedical [1], etc.)[2], and major companies are employing RDF to allow users to mark up their data (e.g., Open Graph Protocol[3], schema.org[4], Twitter cards[5]).

Thus there is an abundance of RDF data which is driving the kinds of data integration challenges we posit FBMs to be valuable for. Even those which use different ontologies can be meaningfully unified using a relatively simple "upper ontology" given a basic knowledge of common ontologies [19]. For non-RDF data sources, the heterogeneity problem can be solved by providing an RDF or SPARQL [15] interface on data sources (either on the producer or consumer side, like in [17], which is a coding task) and by providing an appropriate ontology to give a unified view of the data (which is a design task needing sufficiently knowledgeable persons or good documentation of the data source).

We experimented with this streaming FBM method on a large RDF dataset crawled from the web. The 2012 Billion Triple Challenge dataset (BTC)[6] is a set of RDF quads crawled from the Web for the purposes of challenging competitors to work at scale. BTC was chosen because it represents one of the best and largest publicly available RDF datasets, and because of our own past experience working with previous versions of it [11], [18].

The RDF quads in BTC are RDF triples with an additional component that we will refer to as the "graph name". For an RDF quad $\langle s, p, o, g\rangle$ (where g is the graph name), let $d = \text{http}(g)$ be the direct URL of the document retrieved over HTTP when following g (e.g., when you put g in your browser). In some cases, $d = g$. However, http(g) can result in redirection (e.g., HTTP codes 301, 302, 303) in which case $d \neq g$. Many graph names can map to the same document URL, and these mappings are also captured in the (broader sense of the) BTC dataset.

For our experiment, we limited ourselves to quads in BTC that utilized terms from the friends-of-a-friend (FOAF) ontology[7]. FOAF contains 164.3M overall, including two specific sub-groups (`foaf:knows` and `foaf:name`) which we will use below (see Table I).

[2]http://www.semantic-web-journal.net/accepted-datasets – last accessed August 8, 2013 – contains an entire list of such datasets.

[3]http://ogp.me/ – last accessed August 8, 2013

[4]http://schema.org/ – last accessed August 8, 2013

[5]https://dev.twitter.com/docs/cards – last accessed August 8, 2013

[6]http://km.aifb.kit.edu/projects/btc-2012/ – last accessed August 8, 2013

[7]http://xmlns.com/foaf/spec/ – last accessed August 8, 2013

FOAF captures information about people, webpages, and their relationships. We considered documents as sources for determining beliefs. Relating to the judge example, we are the judge, and each document is a witness. We assume every document asserts that it is telling the absolute truth for each triple. That is, for each $\langle s, p, o, g\rangle$ such that $d = \text{http}(g)$, d's belief of $\langle s, p, o\rangle$ is $\langle 1, 0, 0\rangle$ (absolute belief). As the judge, we must determine how to discount these beliefs since we are not inclined to believe anything absolutely just by mere testimony.

Hogan *et al.* [9] have already established a notion of authority for RDF triples based on the sources of the triples, and so we make use of that. It is important to understand that whether an RDF triple is authoritative depends on the relationship between the subject of the RDF triple (that is, s in $\langle s, p, o\rangle$) and the graph name g and/or document URL d. The general idea is that any RDF triple in which the subject is a term defined by the source, such an RDF triple is considered authoritative. The foundation for this notion is laid by concepts of RDF namespaces and Linked Data principles[8], the scope of which are beyond this particular work.

The specific rules we used are as follows.

- For any URI u, let nofrag(u) be everything before the fragment (if any fragment exists). Thus nofrag(data:abc) = data:abc, and nofrag(data:abc#def) = data:abc (the "fragment" is everything after and including the first "#" in a URI). $\langle s, p, o, g\rangle$ is considered authoritative iff nofrag(s)=nofrag(g) or nofrag(s)=nofrag(http(g)).
- We transform $\langle s, p, o, g\rangle$ RDF quads into quints $\langle s, p, o, d, a\rangle$ where $d = \text{http}(g)$ and $a = 1$ if $\langle s, p, o, g\rangle$ is authoritative and $a = 0$ otherwise, and we consider only unique quints. If $a = 1$, our belief in d for the assertion of $\langle s, p, o\rangle$ is $\langle 0.9, 0, 0.1\rangle$ (90% belief), and if $a = 0$, our belief in d for the assertion of $\langle s, p, o\rangle$ is $\langle 0.01, 0, 0.99\rangle$ (99% uncertainty). The values are chosen somewhat arbitrarily. Since d's belief of $\langle s, p, o\rangle$ is always $\langle 1, 0, 0\rangle$ and $X \otimes \langle 1, 0, 0\rangle = X$, our belief in d becomes our belief of d's assertion of $\langle s, p, o\rangle$. (That is, unsurprisingly, our belief in the source for a particular triple is the same as our belief in the triple stated by that source.)
- At this point, we have beliefs for every unique $\langle s, p, o, d, a\rangle$, which we will denote $b(\langle s, p, o, d, a\rangle)$, but we wish to form some overall belief for each unique $\langle s, p, o\rangle$. This is determined using the consensus operator $\oplus$. For every $\langle s, p, o\rangle$, our belief is

$$b(\langle s, p, o\rangle) = \bigoplus_{\langle s,p,o,d,a\rangle \in BTC} b(\langle s, p, o, d, a\rangle).$$

C. Implementation

Our evaluation was run using simple Unix commands like `sort`, `cut`, and `uniq`; and short, custom Perl scripts. This was simply the easiest path to a preliminary evaluation of FBMs. In principle, though, the same computation is easily parallelizable. Let S be a set of data sources (documents),

[8]http://www.w3.org/DesignIssues/LinkedData.html – last accessed August 8, 2013

and let W be a function associating our opinions *about* data sources in S to those same data sources. Let K (the "knowledge base") be a set of opinions *from* the sources in S. Then generalizing the previous formula for BTC, we form our opinion $w(R)$ of a proposition R as:

$$w(R) = \bigoplus_{w_A(R) \in K} [W(A) \otimes w_A(R)].$$

To parallelize this computation, simply artibrarily distribute K to some n processors, that is, $K = \bigcup_{i=0}^{n-1} K_i$. Then each processor can determine its *local* consensus opinions K_i' as:

$$K_i' = \left\{ \bigoplus_{w_A(R) \in K_i} \left[W(A) \otimes w_A(R)\right] \;\middle|\; \forall R \right\}.$$

Then in order to derive the *global* consensus opinions, a simple parallel reduction using the $\oplus$ operator is possible by virtue of the fact that $\oplus$ is associative and commutative. For example, each processor i may have a *local* consensus opinion $w_i(R)$. Then the *global* consensus opinion of R is $\bigoplus_{i=0}^{n-1} w_i(R)$, derived using a parallel reduction. Alternatively, for every proposition R, a hash function h can be used to redistribute *local* consensus opinions among processors so that processor $h(R) \bmod n$ has all of $\{w_i(R)\}_{i=0}^{n-1}$ and the derivation of the *global* opinions can be performed as local operations.

D. Results

The distribution of the consensus beliefs are illustrated in Fig. 5. Each $\langle X, Y \rangle$ point is the number Y of unique $\langle s, p, o \rangle$ triples for which our belief is X. The red points represent authoritative triples, and the blue points represent non-authoritative triples. For a closer view of the highest beliefs, refer to Fig. 6.

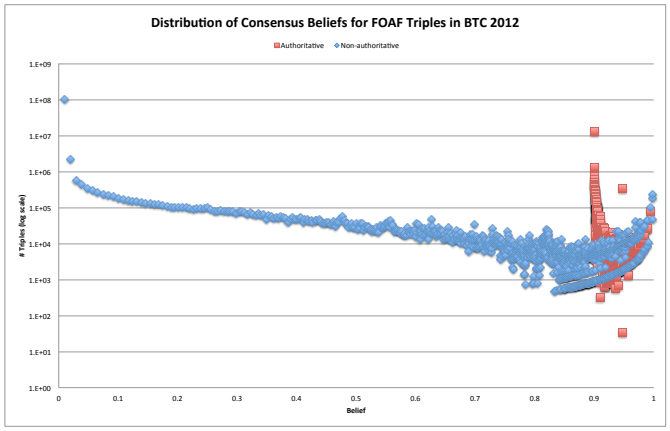

Fig. 5. Each $\langle X, Y \rangle$ point is the number Y of unique $\langle s, p, o \rangle$ triples for which our belief is X. The red points represent authoritative triples, and the blue points represent non-authoritative triples.

We note a good distribution over the range of belief values (i.e., there are no significant gaps across the horizontal axis) which suggests that belief as specified herein provides a useful ranking mechanism. Second, there are a large number of triples

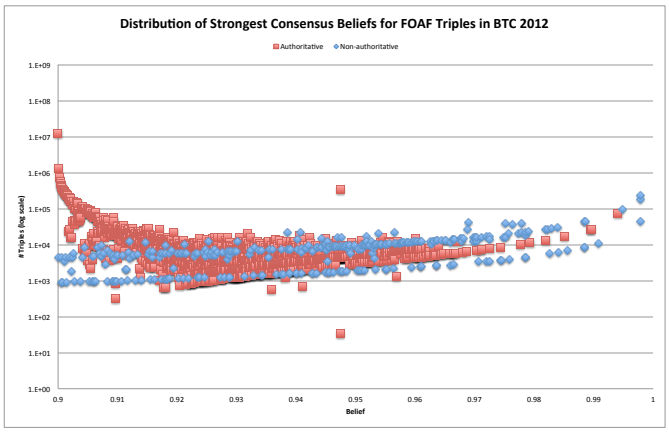

Fig. 6. Each $\langle X, Y \rangle$ point is the number Y of unique $\langle s, p, o \rangle$ triples for which our belief is X. The red points represent authoritative triples, and the blue points represent non-authoritative triples.

for which we have very little belief and a large number of triples for which we have high belief, which indicates that belief as specified herein is a useful metric for separating highly believable statements from hardly believable statements (as long as the underlying assumptions of authority hold and are meaningful in reality). Third, our most believed triples are actually non-authoritative, reflecting strong public consensus about these triples/propositions even without an authoritative source.

Diving deeper into the data, it appears that the overall shape of the charts is caused (at least in part) by distribution of names, as shown in figure 7. It so happens that documents often include the names of people mentioned even if the document is not authoritative for that person. For example, Jesse may have a document that is authoritative for statements about j:jesse, and so the document may say that ⟨j:jesse, foaf:knows, c:cliff⟩ and also that ⟨c:cliff, foaf:name, "Cliff"⟩, even though the document is only authoritative for j:jesse and not c:cliff. We conjecture that popular persons have their names replicated across relatively non-authoritative documents which accounts for the high belief in some non-authoritative triples.

If we look at only triples with foaf:knows as a predicate, the disparity in Fig. 8 is quite obvious. Non-authoritative foaf:knows triples are hardly believed while authoritative foaf:knows triples are very believed. We conjecture that this is because the publication behavior of foaf:knows triples is opposite to that of foaf:name triples. For example, Jesse's document may state that ⟨j:jesse, foaf:knows, c:cliff⟩, but it does not state that ⟨c:cliff, foaf:knows, j:jesse⟩. Clearly, such triples of the latter case exist or else no non-authoritative foaf:knows triples would even exist, but such triples are uncommon which leads to low belief.

IV. Conclusion

This work represents the beginning of our investigation, and indeed more is necessary to verify our conjectures and to

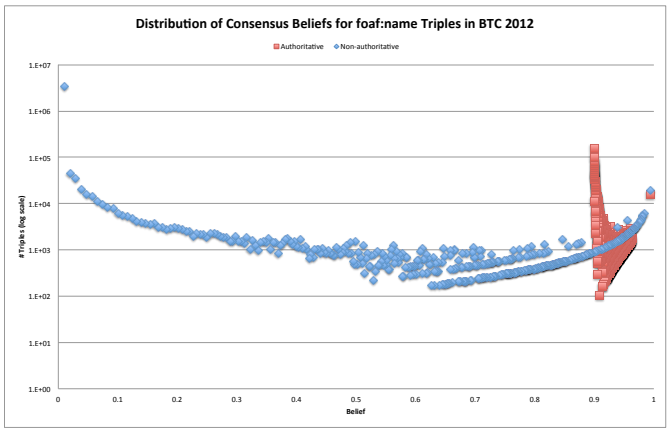

Fig. 7. Belief distribution for the 7.8M triples with predicate foaf:name.

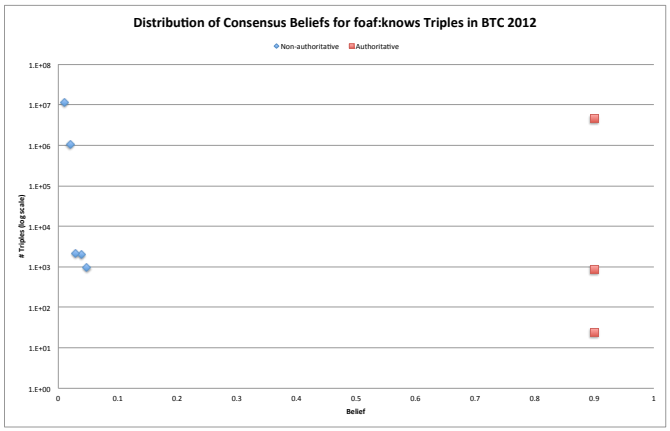

Fig. 8. Belief distribution for the 17.3M triples with predicate foaf:knows in BTC.

find more patterns. Regardless, this preliminary work indicates that focused belief measures hold promise and that more investigation is warranted.

The most significant issue is in our use of "dogmatic" base claims, that is, opinions of the form $\langle 1, 0, 0\rangle$ or $\langle 0, 1, 0\rangle$, expressing complete belief or disbelief on the part of the claimant. In fact, truth claims come in all forms, e.g. "A believes that p" or "A holds p with 50% probability" or "A believes that p falls in the range $[10, 20]$", and many other possible forms involving intervals, distributions, statistical properties, etc. Being able to map these source claims to FBM opinions is an important next problem for us.

Other future work includes:

- The current experiments depend on a number of constant assumptions, as shown above. Parameterization of these constants will support a sensitivity analysis over this space of inputs to help determine experimental behavior.
- Discovering more complex categorizations of triples on the web than merely "authoritative" and "non-authoritative" and determining meaningful discounting opinions for these categorizations.
- Taking into account negative assertions (that is, asserting the falsity of a triple). Such is supported by OWL, but it is expressed using multiple triples. Our evaluation herein equated each single triple as a single proposition.
- Implementation of a parallel system for deriving consensus opinions as described in section III-C.

References

[1] Michael Bada, Kevin Livingston, and Lawrence Hunter. An ontology representation of biomedical data sources and records. *Bio-Ontologies 2011*, 2011.

[2] Jeremy J. Carroll and Graham Klyne. Resource description framework (RDF): Concepts and abstract syntax. W3C recommendation, W3C, February 2004. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.

[3] Brian F Chellas. *Modal logic*. Cambridge university press, 1980.

[4] Li Ding, Dominic DiFranzo, Alvaro Graves, James R. Michaelis, Xian Li, Deborah L. McGuinness, and James A. Hendler. TWC data-gov corpus: incrementally generating linked government data from data.gov. In *Proceedings of the 19th International Conference on the World Wide Web*, 2010.

[5] Pedro Domingos and Daniel Lowd. *Markov Logic: An Interface Layer for Artificial Intelligence*. Morgan and Claypool, 2009.

[6] Michel Grabisch. Belief functions on lattices. *Int. J. Intelligent Systems*, 24:1:76–95, 2009.

[7] W3C OWL Working Group. OWL 2 web ontology language document overview. Technical report, W3C, December 2012. http://www.w3.org/TR/2012/REC-owl2-overview-20121211/.

[8] Patrick Hayes. RDF semantics. W3C recommendation, W3C, February 2004. http://www.w3.org/TR/2004/REC-rdf-mt-20040210/.

[9] Aidan Hogan, Jeff Z. Pan, Axel Polleres, and Stefan Decker. Saor: template rule optimisations for distributed reasoning over 1 billion linked data triples. In *Proceedings of the 9th international semantic web conference*, pages 337–353, 2010.

[10] Audun Josang. A logic for uncertain probabilities. *Int. J. Uncertainty, Fuzziness and Knowledge-Based Systems*, 9:3:279–311, 2001.

[11] Cliff Joslyn, Bob Adolf, Sinan al Saffar, J Feo, Eric Goodman, David Haglin, Greg Mackey, and David Mizell. High performance descriptive semantic analysis of semantic graph databases. In *Proc. Wshop. on High Performance Computing for the Semantic Web (HPCSW 2011), CEUR*, volume 736, 2011.

[12] Cliff Joslyn and Jane Booker. Generalized information theory for engineering modeling and simulation. In E Nikolaidis et al., editor, *Engineering Design Reliability Handbook*, pages 9:1–40. CRC Press, 2005.

[13] Nils J Nilsson. Probabilistic logic. *Artificial intelligence*, 28(1):71–87, 1986.

[14] Donald Nute. Defeasible logic. In Oskar Bartenstein, Ulrich Geske, Markus Hannebauer, and Osamu Yoshie, editors, *Web Knowledge Management and Decision Support*, volume 2543 of *Lecture Notes in Computer Science*, pages 151–169. Springer Berlin Heidelberg, 2003.

[15] Eric Prud'hommeaux and Andy Seaborne. SPARQL query language for RDF. W3C recommendation, W3C, January 2008. http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/.

[16] P Walley. Towards a unified theory of imprecise probabilities. *Int. J. Approximate Reasoning*, 24:125–148, 2000.

[17] Jesse Weaver and Paul Tarjan. Facebook Linked Data via the Graph API. *Semantic Web Journal*, 2012.

[18] Gregory T Williams, Jesse Weaver, Medha Atre, and James A Hendler. Scalable reduction of large datasets to interesting subsets. In *Billion Triple Challenge, ISWC 2009*, 2009.

[19] Gregory Todd Williams, Jesse Weaver, Medha Atre, and James A Hendler. Scalable reduction of large datasets to interesting subsets. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(4):365–373, 2010.