# On Bayesian Interpretation of Fact-finding in Information Networks

Dong Wang, Tarek Abdelzaher, Hossein Ahmadi, Jeff Pasternack,
Dan Roth, Manish Gupta, Jiawei Han, Omid Fatemieh, and Hieu Le
Department of Computer Science, University of Illinois
Urbana, IL 61801

Charu C. Aggarwal
IBM Research
Yorktown Heights, N.Y. 10598

*Abstract*—When information sources are unreliable, information networks have been used in data mining literature to uncover facts from large numbers of complex relations between noisy variables. The approach relies on topology analysis of graphs, where nodes represent pieces of (unreliable) information and links represent abstract relations. Such topology analysis was often empirically shown to be quite powerful in extracting useful conclusions from large amounts of poor-quality information. However, no systematic analysis was proposed for quantifying the accuracy of such conclusions. In this paper, we present, for the first time, a Bayesian interpretation of the basic mechanism used in fact-finding from information networks. This interpretation leads to a direct quantification of the accuracy of conclusions obtained from information network analysis. Hence, we provide a general foundation for using information network analysis not only to heuristically extract likely facts, but also to quantify, in an analytically-founded manner, the probability that each fact or source is correct. Such probability constitutes a measure of quality of information (QoI). Hence, the paper presents a new foundation for QoI analysis in information networks, that is of great value in deriving information from unreliable sources. The framework is applied to a representative fact-finding problem, and is validated by extensive simulation where analysis shows significant improvement over past work and great correspondence with ground truth.

**Keywords: Information networks, sensors, Bayesian inference.**

## I. INTRODUCTION

Information networks are a key abstraction in data mining literature used to uncover facts from a large number of relations between unreliable observations [1]. The power of information network analysis lies in its ability to extract useful conclusions even when the degree of reliability of the input data or observations is not known in advance. For example, given a set of claims from a multitude of sources, one can rank both the claimed information pieces (let us call them *assertions*) and their sources by credibility, given no *a priori* knowledge of the truthfulness of the individual assertions and sources. Alternatively, given only data on who publishes in which conferences one can rank both the authors and the conferences by authority in the field.

This paper presents a new analytic framework that enables, for the first time, the calculation of correct probabilities of conclusions resulting from information network analysis. Such probabilities constitute a measure of quality of information (QoI). Our analysis relies on a Bayesian interpretation of the basic inference mechanism used for fact-finding in information network literature.

In the simplest version of fact-finding from information networks, nodes represent entities such as sources and assertions. Edges denote their relations (e.g., who claimed what). Each category of nodes is then iteratively ranked. Assertions are given a ranking that is proportional to the number of their sources, each source weighted by its credibility. Sources are then given a ranking that is proportional to the number of the assertions they made, each weighted by its credibility. This iterative ranking process continues until it converges. Information network analysis is good at such ranking. While the algorithms compute an intuitive "credibility score", as we demonstrate in this paper, they do not actually compute the real *probability* that a particular conclusion is true. For example, given that some source is ranked 17th by credibility, it is not clear what that means in terms of probability that the source says the truth. Our paper addresses this problem, providing a general analytic foundation for quantifying the probability of correctness in fact-finding literature. We show that the probabilities computed using our analysis are significantly more accurate than prior work.

The fact-finding techniques addressed in the paper are particularly useful in environments where a large number of sources are used whose reliability is not *a priori* known (as opposed collecting information from a small number of well-characterized sources). Such situations are common when, for instance, crowd-sourcing is used to obtain information, or when information is to be gleaned from informal sources such as Twitter messages. We focus on networks of sources and assertions. The Bayesian interpretation derived in this paper allows us to accurately quantify the probability that a source is truthful or that an assertion is true in the absence of detailed prior knowledge. Note that, while only source/assertion networks are considered, the analysis allows us to represent a much broader category of information networks. For example, in the author/conference network, one can interpret the act of publishing in a conference as an implicit assertion that the conference is good. The credibility of the assertion depends on the authority of the author. Hence, the network fits the source/assertion model.

This paper is intended to be a first step towards a new category of information network analysis. Being the first step,

we focus on laying the foundations, such that extensions of this work can easily adapt the analysis to more complex information network models. Hence, we start with a very simple model, leaving extensions to future work.

With that in mind, in Section II of this paper, we present a Bayesian analysis of source/assertion information networks. In Section III, we evaluate the results using extensive simulations involving examples of up to 100 sources and 1000s of assertions. Section IV discusses limitations of the approach. Section V presents related work. The paper concludes with Section VI.

## II. FACT-FINDING IN INFORMATION NETWORKS

This paper presents a foundation for quality of information analysis in information networks. We are specifically interested in the network model used for deriving credibility of facts and sources. We call the iterative ranking algorithm used for analyzing source/assertion information networks, a *fact-finder*. The algorithm ranks a list of assertions and a list of sources by credibility. In the following subsections, we first review the basic algorithm, then propose its Bayesian interpretation that allows quantifying the actual probability that a source is truthful or that an assertion is true.

### A. The Basic Fact-finder

Let there be $s$ sources, $S_1, ..., S_s$ who collectively assert $c$ different pieces of information, $C_1, ..., C_c$. We call each such piece of information an assertion. We represent all sources and assertions by a network, where these sources and assertions are nodes, and where a claim, $C_{i,j}$ (denoting that a source $S_i$ makes assertion $C_j$) is represented by a link between the corresponding source and assertion nodes. We assume that a claim can either be true or false. An example is "John Smith is CEO of Company X" or "Building Y is on Fire". We further define $Cred(S_i)$ as the credibility of source $S_i$, and $Cred(C_j)$ as the credibility of assertion $C_j$.

Algebraically, we define the $c \times 1$ vector, $\overline{C}_{cred}$, to be the assertion credibility vector $[Cred(C_1)...Cred(C_c)]^T$ and the $s \times 1$ vector, $\overline{S}_{cred}$, to be the source credibility vector $[Cred(S_1)...Cred(S_s)]^T$. We also define the $c \times s$ array $CS$ such that element $CS(j,i) = 1$ if source $S_i$ makes claim $C_j$, and is zero otherwise.

Now let us define $\overline{C}_{cred}^{est}$ as a vector of *estimated* assertion credibility, defined as $(1/\alpha)[CS]\overline{S}_{cred}$. One can pose the basic fact-finding problem as one of finding a least squares estimator (that minimizes the sum of squares of errors in source credibility estimates) for the following system:

$$\overline{C}_{cred}^{est} = \frac{1}{\alpha}[CS]\overline{S}_{cred} \qquad (1)$$

$$\overline{S}_{cred} = \frac{1}{\beta}[CS]^T\overline{C}_{cred}^{est} + \overline{e} \qquad (2)$$

where the notation $X^T$ denotes the transpose of matrix $X$. It can further be shown that the condition for it to minimize the error is that $\alpha$ and $\beta$ be chosen such that their product is an Eigenvalue of $[CS]^T[CS]$. The algorithm produces the credibility values $Cred(S_i)$ and $Cred(C_j)$ for every source $S_i$ and for every assertion $C_j$. These values are used for ranking. The question is, does the solution have an interpretation that allows quantifying the actual probability that a given source is truthful or that a given assertion is true? The question is answered in the next section.

### B. A Bayesian Interpretation

Let $S_i^t$ denote the proposition that "Source $S_i$ speaks the truth". Let $C_j^t$ denote the proposition that "Assertion $C_j$ is true". Also, let $S_i^f$ and $C_j^f$ denote the negation of the above propositions, respectively. Our objective, in this section, is to estimate the probabilities of these propositions. We further define $S_iC_j$ to mean "Source $S_i$ made assertion $C_j$".

It is useful to define $Claims_i$ as the set of all claims made by source $S_i$, and $Sources_j$ as the set of all sources who claimed assertion $C_j$. In the subsections below, we derive the posterior probability that an assertion is true, followed by the derivation of the posterior probability that a source is truthful.

*1) Assertion Credibility:* Consider some assertion $C_j$, claimed by a set of sources $Sources_j$. Let $i_k$ be the $k$th source in $Sources_j$, and let $|Sources_j| = K_j$. (For notational simplicity, we shall occasionally omit the subscript $j$ from $K_j$ in the discussion below, where no ambiguity arises.) According to Bayes theorem:

$$P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) =$$
$$\frac{P(S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j|C_j^t)}{P(S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j)}P(C_j^t) \qquad (3)$$

The above equation makes the implicit assumption that the probability that a source makes any given assertion is sufficiently low that no appreciable change in posterior probability can be derived from the lack of a claim (i.e., lack of an edge between a source and an assertion). Hence, only existence of claims is taken into account. Assuming further that sources are conditionally independent (i.e., given an assertion, the odds that two sources claim it are independent), Equation (3) is rewritten as:

$$P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) =$$
$$\frac{P(S_{i_1}C_j|C_j^t)...P(S_{i_K}C_j|C_j^t)}{P(S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j)}P(C_j^t) \qquad (4)$$

Let us further assume that the change in posterior probability we get from any single source or claim is small. This is typical when using evidence collected from many individually unreliable sources. Hence:

$$\frac{P(S_{i_k}C_j|C_j^t)}{P(S_{i_k}C_j)} = 1 + \delta_{i_kj}^t \qquad (5)$$

where $|\delta_{i_kj}^t| << 1$. Similarly:

$$\frac{P(S_{i_k}C_j|C_j^f)}{P(S_{i_k}C_j)} = 1 + \delta_{i_kj}^f \qquad (6)$$

where $|\delta_{i_k j}^f| << 1$. Under the above assumptions, we prove in Appendix A that the denominator of the right hand side in Equation (4) can be rewritten as follows:

$$P(S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) \approx \prod_{k=1}^{K_j} P(S_{i_k}C_j) \qquad (7)$$

Please see Appendix A for a proof of Equation (7). Note that, the proof does *not* rely on an independence assumption of the marginals, $P(S_{i_k}C_j)$. Those marginals are, in fact, not independent. The proof merely shows that, under the assumptions stated in Equation (5) and Equation (6), the above approximation holds true. Substituting in Equation (4):

$$P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) \quad = \quad \frac{P(S_{i_1}C_j|C_j^t)...P(S_{i_K}C_j|C_j^t)}{P(S_{i_1}C_j)...P(S_{i_K}C_j)} P(C_j^t) \qquad (8)$$

which can be rewritten as:

$$P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) \quad = \quad \frac{P(S_{i_1}C_j|C_j^t)}{P(S_{i_1}C_j)}$$
$$\times \quad ...$$
$$\times \quad \frac{P(S_{i_K}C_j|C_j^t)}{P(S_{i_K}C_j)}$$
$$\times \quad P(C_j^t) \qquad (9)$$

Substituting from Equation (5):

$$P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) \quad = \quad P(C_j^t)\prod_{k=1}^{K_j}(1+\delta_{i_k j}^t)$$
$$= \quad P(C_j^t)(1+\sum_{k=1}^{K_j}\delta_{i_k j}^t) \qquad (10)$$

The last line above is true because higher products of $\delta_{i_k j}^t$ can be neglected, since we assumed $|\delta_{i_k j}^t| << 1$. The above equation can be re-written as:

$$\frac{P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) - P(C_j^t)}{P(C_j^t)} = \sum_{k=1}^{K_j}\delta_{i_k j}^t \quad (11)$$

where, from Equation (5):

$$\delta_{i_k j}^t = \frac{P(S_{i_k}C_j|C_j^t) - P(S_{i_k}C_j)}{P(S_{i_k}C_j)} \qquad (12)$$

*2) Source Credibility:* Next, consider some source $S_i$, who makes the set of claims $Claims_i$. Let $j_k$ be the $k$th claim in $Claims_i$, and let $|Claims_i| = L_i$. (For notational simplicity, we shall occasionally omit the subscript $i$ from $L_i$ in the discussion below, where no ambiguity arises.) According to Bayes theorem:

$$P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L}) \quad = \quad \frac{P(S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L}|S_i^t)}{P(S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L})} P(S_i^t) \qquad (13)$$

As before, assuming conditional independence:

$$P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L}) \quad = \quad \frac{P(S_iC_{j_1}|S_i^t)...P(S_iC_{j_L}|S_i^t)}{P(S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L})} P(S_i^t) \qquad (14)$$

Once more we invoke the assumption that the change in posterior probability caused from any single claim is very small, we get:

$$\frac{P(S_iC_{j_k}|S_i^t)}{P(S_iC_{j_k})} = 1 + \eta_{ij_k}^t \qquad (15)$$

where $|\eta_{ij_k}^t| << 1$. Similarly to the proof in Appendix A, this leads to:

$$P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L}) \quad = \quad \frac{P(S_iC_{j_1}|S_i^t)}{P(S_iC_{j_1})}$$
$$\times \quad ...$$
$$\times \quad \frac{P(S_iC_{j_L}|S_i^t)}{P(S_iC_{j_L})}$$
$$\times \quad P(S_i^t) \qquad (16)$$

We can then re-write Equation (16) as follows:

$$P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L}) \quad = \quad P(S_i^t)\prod_{k=1}^{L_i}(1+\eta_{ij_k}^t)$$
$$= \quad P(S_i^t)(1+\sum_{k=1}^{L_i}\eta_{ij_k}^t) \qquad (17)$$

The above equation can be further re-written as:

$$\frac{P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L}) - P(S_i^t)}{P(S_i^t)} = \sum_{k=1}^{L_i}\eta_{ij_k}^t \quad (18)$$

where, from Equation (15):

$$\eta_{ij_k}^t = \frac{P(S_iC_{j_k}|S_i^t) - P(S_iC_{j_k})}{P(S_iC_{j_k})} \qquad (19)$$

*3) The Iterative Algorithm:* In the sections above, we derived the expressions of posterior probability that an assertion is true or that a source is truthful. These expressions were derived in terms of $\delta_{i_k j}^t$ and $\eta_{ij_k}^t$. It remains to show how these quantities are related. Let us first consider the terms in Equation (12) that defines $\delta_{i_k j}^t$. The first is $P(S_iC_j|C_j^t)$, the probability that $S_i$ claims assertion $C_j$, given that $C_j$ is true. (For notational simplicity, we shall use subscripts $i$ and $j$ to denote the source and the assertion.) We have:

$$P(S_iC_j|C_j^t) = \frac{P(S_iC_j, C_j^t)}{P(C_j^t)} \qquad (20)$$

where:

$$P(S_iC_j, C_j^t) \quad = \quad P(S_i \ speaks)$$
$$P(S_i \ claims \ C_j|S_i \ speaks)$$
$$P(C_j^t|S_i \ speaks, S_i \ claims \ C_j) \qquad (21)$$

In other words, the joint probability that link $S_iC_j$ exists and $C_j$ is true is the product of the probability that $S_i$ speaks, denoted $P(S_i\ speaks)$, the probability that it claims $C_j$ given that it speaks, denoted $P(S_i\ claims\ C_j|S_i\ speaks)$, and the probability that the assertion is true, given that it is claimed by $S_i$, denoted $P(C_j^t|S_i\ speaks, S_i\ claims\ C_j)$. Here, $P(S_i\ speaks)$ depends on the rate at which $S_i$ makes assertions. Some sources may be more prolific than others. $P(S_i\ claims\ C_j|S_i\ speaks)$ is simply $1/c$, where $c$ is the total number of assertions. Finally, $P(C_j^t|S_i\ speaks, S_i\ claims\ C_j)$ is the probability that $S_i$ is truthful. Since we do not know ground truth, we estimate that probability by the best information we have, which is $P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L})$. Thus:

$$P(S_iC_j, C_j^t) = \frac{P(S_i\ speaks)P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L})}{c} \quad (22)$$

Substituting in Equation (20) from Equation (22) and noting that $P(C_j^t)$ is simply the ratio of true assertions, $c_{true}$ to the total assertions, $c$, we get:

$$P(S_iC_j|C_j^t) = \frac{P(S_i\ speaks)P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L})}{c_{true}} \quad (23)$$

Similarly,

$$P(S_iC_j) = \frac{P(S_i\ speaks)}{c} \quad (24)$$

Substituting from Equation (23) and Equation (24) into Equation (12) and re-arranging, we get:

$$\begin{aligned} \delta_{i_kj}^t &= \frac{P(S_{i_k}C_j|C_j^t) - P(S_{i_k}C_j)}{P(S_{i_k}C_j)} \\ &= \frac{P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L})}{c_{true}/c} - 1 \quad (25) \end{aligned}$$

If we take the fraction of all true assertions to the total number of assertions as the prior probability that a source is truthful, $P(S_i^t)$ (which is a reasonable initial guess in the absence of further evidence), then the above equation can be re-written as:

$$\delta_{i_kj}^t = \frac{P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L})}{P(S_i^t)} - 1 \quad (26)$$

Substituting for $\delta_{i_kj}^t$ in Equation (11), we get:

$$\frac{P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) - P(C_j^t)}{P(C_j^t)} =$$

$$\sum_{i=1}^{K_j} \frac{P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L}) - P(S_i^t)}{P(S_i^t)} \quad (27)$$

We can similarly prove that:

$$\eta_{ij_k}^t = \frac{P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j)}{P(C_j^t)} - 1 \quad (28)$$

and:

$$\frac{P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L}) - P(S_i^t)}{P(S_i^t)} =$$

$$\sum_{j=1}^{L_i} \frac{P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) - P(C_j^t)}{P(C_j^t)} \quad (29)$$

Comparing the above equations to the iterative formulation of the basic fact-finder, described in Section II-A, we arrive at the sought interpretation of the credibility rank of sources $Rank(S_i)$ and credibility rank of assertions $Rank(C_j)$ in iterative fact-finding. Namely:

$$Rank(C_j) = \frac{P(C_j^t|S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) - P(C_j^t)}{P(C_j^t)} \quad (30)$$

$$Rank(S_i) = \frac{P(S_i^t|S_iC_{j_1}, S_iC_{j_2}, ..., S_iC_{j_L}) - P(S_i^t)}{P(S_i^t)} \quad (31)$$

In other words, $Rank(C_j)$ is interpreted as the increase in the posterior probability that an assertion is true, normalized by the prior. Similarly, $Rank(S_i)$ is interpreted as the increase in the posterior probability that a source is truthful, normalized by the prior. Substituting from Equation (30) and Equation (31) into Equation (27) and Equation (29), we then get:

$$\begin{aligned} Rank(C_j) &= \sum_{k \in Sources_j} Rank(S_k) \\ Rank(S_i) &= \sum_{k \in Claims_i} Rank(C_k) \quad (32) \end{aligned}$$

Once the credibility ranks are computed such that they satisfy the above equations (and any other problem constraints), Equation (30) and Equation (31), together with the assumption that prior probability that an assertion is true is initialized to $p_a^t = c_{true}/c$, give us the main contribution of this paper, Namely[1]:

$$P(C_j^t|network) = p_a^t(Rank(C_j) + 1) \quad (33)$$

We can similarly show that if $p_s^t$ is the prior probability that a randomly chosen source tells the truth, then:

$$P(S_i^t|network) = p_s^t(Rank(S_i) + 1) \quad (34)$$

Hence, the above Bayesian analysis presents, for the first time, a basis for estimating the probability that *each individual source*, $S_i$, is truthful and that *each individual assertion*, $C_j$, is true. These two vectors are computed based on two scalar constants: $p_a^t$ and $p_s^t$, which represent estimated statistical averages over all assertions and all sources, respectively.

## III. EVALUATION

In this section, we carry out experiments to verify the correctness and accuracy of the probability that a source is truthful or an assertion is true predicted from the Bayesian interpretation of fact-finding in information networks. We then compare our techniques to previous algorithms in fact-finder literature.

We built a simulator in Matlab 7.8.0 to simulate the source and assertion information network. To test our results, we

---

[1]The equations above are ambiguous with respect to a scale factor. To handle the ambiguity we impose the constraint that probabilities cannot exceed one.

generate a random number of sources and assertions, and partition these assertions into true and false ones. A random probability, $P_i$, is assigned to each source $S_i$ representing the ground truth probability that the source speaks the truth. For each source $S_i$, we then generate $L_i$ claims. Each claim has a probability $P_i$ of being true and a probability $1 - P_i$ of being false. A true claim links the source to a randomly chosen true assertion (representing that the source made that assertion). A false claim links the source to a randomly chosen false assertion. This generates an information network.

We let $P_i$ be uniformly distributed between 0.5 and 1 in our experiments[2]. We then find an assignment of credibility values that satisfies Equation (32) for the topology of the generated information network. Finally, we compute the estimated probability that an assertion is true or a source is truthful from the resulting credibility values of assertions and sources based on Equation (33) and (34). Since we assumed that claims are either true or false, we view each assertion as "true" or "false" based on whether the probability that it is true is above or below 50%. Then the computed results are compared against the ground truth to report the prediction accuracy.

For sources, we simply compare the computed probability to the ground truth probability that they tell the truth. For assertions, we define two metrics to evaluate prediction accuracy: false positives and false negatives. The false positives are defined as the ratio of the number of false assertions that are classified as true over the total number of assertions that are classified as true. The false negatives are defined as the ratio of the number of true assertions that are classified as false over the total number of assertions that are classified as false. For each given source correctness probability (i.e., ground truth) distribution, we average the results over 100 network topologies (e.g., datasets over a time series). Reported results are averaged over 100 random source correctness probability distributions.

In the first experiment, we show the effect of the number of sources on prediction accuracy. We fix the number of true and false assertions at 1000 respectively. We set the average number of claims per source to 100. The number of sources is varied from 10 to 100. The prediction accuracy for both sources and assertions is shown in Figure 1. We note that both false positives and false negatives decrease as the number of sources grows. For more than 40 sources less than 1% of assertions are misclassified. The source correctness probability prediction exhibits a relatively small error (between 3% and 6%). The error first increases and then decreases as the number of sources increases. The reason is that there are two conflicting factors that affect the credibility prediction accuracy of sources: i) average number of assertions per source and ii) average number of sources per assertion. As the the number of sources increases, the first factor decreases (reduce source credibility prediction accuracy) and the second factor increases (improve assertion and eventually source credibility

prediction accuracy). When the number of sources is small, the change of the first factor is more significant than the second, thus its effect dominates. As the number of sources increases, the effect of the second factor overweights the first one and makes source correctness probability prediction error reduce.

Note that, the source correctness probability prediction is especially accurate (e.g., error is around 0.03) when the number of sources is relatively large. At the same time, both the false positives and false negatives in assertion classification are near zero under those conditions, illustrating that the approach has good scalability properties. Its usefulness increases for large networks.

The next experiment shows the effect of changing the assertion mix on prediction accuracy. We vary the ratio of the number of true assertions to the total number of assertions in the network. Assuming that there is usually only one variant of the truth, whereas rumors have more versions, one might expect the set of true assertions to be smaller than the set of false ones. Hence, we fix the total number of assertions to be 2000 and change the ratio of true to total assertions from 0.1 to 0.6. The number of sources in the network is set to 30. The prediction accuracy for both sources and assertions is shown in Figure 2. Observe that the source correctness probability prediction error decreases as the ratio of true assertions increases. This is intuitive: more independent true assertions can be used to improve credibility estimates of sources. Additionally, the false positives and false negatives increase because the true assertion set becomes less densely claimed and more true and false assertions are misclassified as each other as the number of true assertions grows.

Finally, we compared our proposed Bayesian interpretation scheme to four other fact-finder schemes: Average-Log [2], Sums(Hubs and Authorities) [3], an adapted PageRank [4] where claims are bidirectional "links" between source and asserted "documents", and TruthFinder [5]. We selected these because, unlike other state-of-art fact-finders (e.g., 3-Estimates [6]), these do not require knowing what mutual exclusion, if any, exists among the assertions. In this experiment, the number of true and false assertions is 1000 respectively, the number of claims per source is 100, and the number of sources varies from 10 to 100. The fact-finder baselines treat multiple source-assertion network topologies as a whole data set. Using the initial assertion beliefs suggested by [2], we ran each baseline fact-finder for 20 iterations, and then selected the 1000 highest-belief assertions as those predicted to be correct. The estimated probability of each source making a true claim was thus calculated as the proportion of predicted-correct claims asserted relative to the total number of claims asserted by source.

The compared results are shown in Figure 3. Observe that the prediction error of source correctness probability by the Bayesian interpretation scheme is significantly lower than all baseline fact-finder schemes. The reason is that Bayesian analysis estimates the source correctness probability more accurately based on Equation (34) derived in the paper rather than using heuristic methods adopted by the baseline schemes.

---

[2]In principle, there is no incentive for a source to lie more than 50% of the time, since negating their statements would then give a more accurate truth
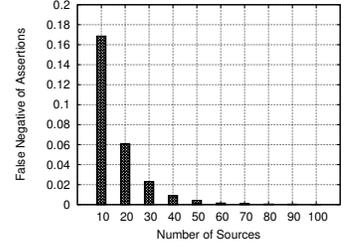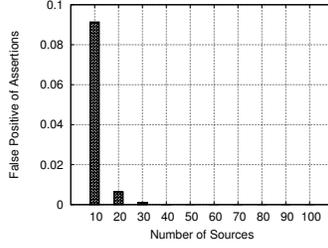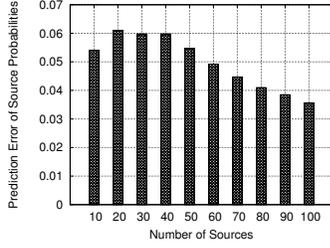
(a) Source Prediction Accuracy     (b) Assertion Prediction False Positives     (c) Assertion Prediction False Negatives

Figure 1.    Prediction Accuracy vs Varying Number of Sources
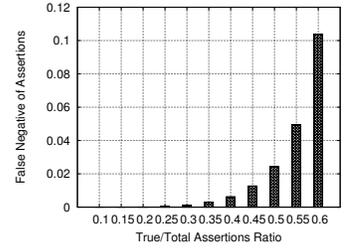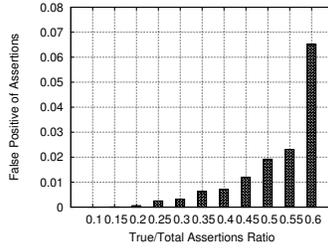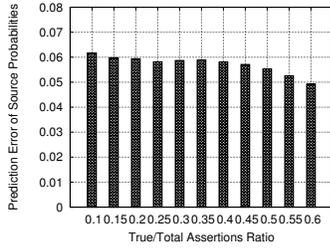


(a) Source Prediction Accuracy     (b) Assertion Prediction False Positives     (c) Assertion Prediction False Negatives

Figure 2.    Prediction Accuracy vs Varying True/Total Assertions
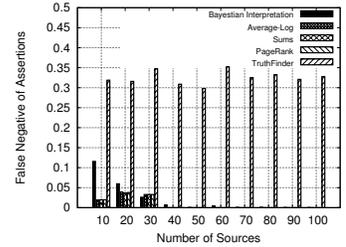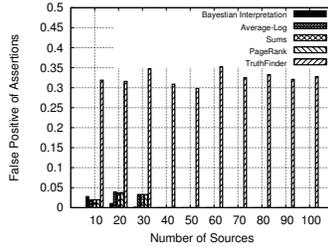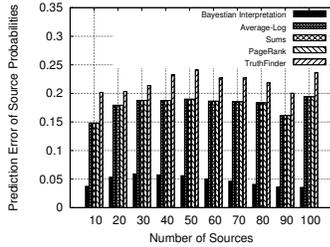


(a) Source Prediction Accuracy     (b) Assertion Prediction False Positives     (c) Assertion Prediction False Negatives

Figure 3.    Prediction Accuracy Comparison with Baseline Fact-finders

We also note that the prediction performance for assertions in the Bayesian scheme is generally as good as the baselines. This is good since the other techniques excel at ranking, which (together with the hint on the number of correct assertions) is sufficient to identify which ones these are. The results confirm the advantages of the Bayesian approach over previous ranking-based work at what the Bayesian analysis does best: estimation of probabilities of conclusions from observed evidence.

## IV. DISCUSSION AND ASSUMPTIONS

This paper presented a Bayesian interpretation of the most basic fact-finding algorithm. The question was to understand why the algorithm is successful at ranking, and to use that understanding to translate the ranking into actual probabilities. Several simplifying assumptions were made that offer opportunities for future extensions.

No dependencies were assumed among different sources or different claims. In reality, sources could be influenced by other sources. Claims could fall into mutual exclusion sets, such as when one is the negation of the other. Taking such relations into account can further improve quality of fact-finding. The change in posterior probabilities due to any single edge in the source-assertion network was assumed to be very small. In other words, we assumed that $|\delta_{i_k j}^t| \ll 1$ and $|\eta_{i j_k}^t| \ll 1$. It is interesting to extend the scheme to situations where a mix of reliable and unreliable sources is used. In this case, assertions from reliable sources can help improve the determination of credibility of other sources.

The probability that any one source makes any one assertion was assumed to be low. Hence, the lack of an edge between a source and an assertion did not offer useful information. There may be cases, however, when the absence of a link between a source and an assertion is important. For example, when a

source is expected to bear on an issue, a source that "withholds the truth" exhibits absence of a link that needs to be accounted for.

Having the basic Bayesian interpretation in place, we can relax the above assumptions and accommodate the mentioned extensions in future work. While the iterative relations may become more cumbersome, we hope that the general outline of the proof can still be followed for these more complex scenarios. The authors are currently pursuing the above extensions.

## V. RELATED WORK

Recently there has been work in the data mining community on performing trust analysis based on the data provided by these sources for different objects. Yin et al. [5] introduced a hueristic-based fact finder algorithm *TruthFinder* which performed trust analysis on a providers-facts network. This work was followed by some more basic fact finder algorithms: *Sums, Average.Log, Investment, Pooled Investment* by Pasternack et al. [2].

Further, some extensions to these basic fact finders have been proposed in the literature. Galland et al. [6] incorporate the notion of hardness of facts in trust analysis by proposing thier algorithms: *Cosine, 2-Estimates, 3-Estimates*. Gupta et al. [7] study trust analysis from a cluster-based perspective. Pasternack et al. [2] incorporate common-sense reasoning into fact finding by updating the confidence scores associated with facts during each iteration of basic fact finder algorithms.

An important factor in trust analysis has been source dependency detection. The Data Management Department at AT&T Labs presented a set of research problems, that deal with detecting source dependency issues, at CIDR [8] and then followed up with detailed problem descriptions and solutions in [9], [10]. A follow-up work from Blanco et al. [11] focuses on copy detection using knowledge from multiple attributes.

Recently, there has been work which performs trust analysis on a homogeneous network of information providers [12] or on a homogeneous network of claims [13]. Note that the work by Yin et al. [13] is the only semi-supervised work on trust analysis as yet.

None of these works above has provided a systematic way to quantify the accuracy of the information concluded from their analysis. In contrast, to the best of our knowledge, we are the first to present a Bayesian interpretation of the basic scheme used in fact-finding from information network to directly quantify the probability that each fact or source is correct in an analytically-founded manner.

## VI. CONCLUSIONS

In this paper, we presented a novel analysis technique for information networks that uses a Bayesian interpretation of the network to assess the credibility of facts and sources. Prior literature that uses information network analysis for fact-finding aims at computing the credibility *rank* of different facts and sources. This paper, in contrast, proposes an analytically founded technique to convert rank to a probability that a fact

is true or that a source is truthful. This paper therefore lays out a foundation for quality of information assurances in iterative fact-finding, a common branch of techniques used in data mining literature for analysis of information networks. The fact-finding techniques addressed in this paper are particularly useful in environments where a large number of sources are used whose reliability is not *a priori* known (as opposed collecting information from a small number of well-characterized sources). Such situations are common when, for instance, crowd-sourcing is used to obtain information, or when information is to be gleaned from informal sources such as Twitter messages. The paper shows that accurate information may indeed be obtained regarding facts and sources even when we do not know the credibility of each source in advance, and where individual sources may generally be unreliable.

## APPENDIX A

Consider an assertion $C_j$ made be several sources $S_{i_1}, ..., S_{i_K}$. Let $S_{i_k}C_j$ denote the fact that source $S_{i_k}$ made assertion $C_j$. We further assume that Equation (5) and Equation (6) hold. In other words:

$$\frac{P(S_{i_k}C_j|C_j^t)}{P(S_{i_k}C_j)} = 1 + \delta_{i_k j}^t$$

$$\frac{P(S_{i_k}C_j|C_j^f)}{P(S_{i_k}C_j)} = 1 + \delta_{i_k j}^f$$

where $|\delta_{i_k j}^t| << 1$ and $|\delta_{i_k j}^f| << 1$.

Under these assumptions, we prove that the joint probability $P(S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j)$, denoted for simplicity by $P(Sources_j)$, is equal to the product of marginal probabilities $P(S_{i_1}C_j), ..., P(S_{i_K}C_j)$.

First, note that, by definition:

$$\begin{aligned}
P(Sources_j) &= P(S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j) \\
&= P(S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j|C_j^t)P(C_j^t) \\
&+ P(S_{i_1}C_j, S_{i_2}C_j, ..., S_{i_K}C_j|C_j^f)P(C_j^f)
\end{aligned}$$
(35)

Using the conditional independence assumption, we get:

$$\begin{aligned}
P(Sources_j) &= P(C_j^t)\prod_{k=1}^{K} P(S_{i_k}C_j|C_j^t) \\
&+ P(C_j^f)\prod_{k=1}^{K} P(S_{i_k}C_j|C_j^f)
\end{aligned}$$
(36)

Using Equation (5) and Equation (6), the above can be rewritten as:

$$\begin{aligned}
P(Sources_j) &= P(C_j^t)\prod_{k=1}^{K_j}(1+\delta_{i_k j}^t)\prod_{k=1}^{K_j} P(S_{i_k}C_j) \\
&+ P(C_j^f)\prod_{k=1}^{K_j}(1+\delta_{i_k j}^f)\prod_{k=1}^{K_j} P(S_{i_k}C_j)
\end{aligned}$$
(37)

and since $|\delta^t_{i_k j}| << 1$ and $|\delta^f_{i_k j}| << 1$, any higher-order terms involving them can be ignored. Hence, $\prod_{k=1}^{K_j}(1 + \delta^t_{i_k j}) = 1 + \sum_{k=1}^{K_j} \delta^t_{i_k j}$, which results in:

$$
\begin{aligned}
P(Sources_j) &= P(C^t_j)(1 + \sum_{k=1}^{K_j} \delta^t_{i_k j}) \prod_{k=1}^{K} P(S_{i_k} C_j) \\
&+ P(C^f_j)(1 + \sum_{k=1}^{K_j} \delta^f_{i_k j}) \prod_{k=1}^{K} P(S_{i_k} C_j)
\end{aligned}
\tag{38}
$$

Distributing multiplication over addition in Equation (38), then using the fact that $P(C^t_j) + P(C^f_j) = 1$ and rearranging, we get:

$$
P(Sources_j) = \prod_{k=1}^{K_j} P(S_{i_k} C_j)(1 + Terms_j)
\tag{39}
$$

where:

$$
Terms_j = P(C^t_j) \sum_{k=1}^{K_j} \delta^t_{i_k j} + P(C^f_j) \sum_{k=1}^{K_j} \delta^f_{i_k j}
\tag{40}
$$

Next, it remains to compute $Terms_j$.

Consider $\delta^t_{i_k j}$ as defined in Equation (5). We can rewrite the equation as follows:

$$
\delta^t_{i_k j} = \frac{P(S_{i_k} C_j | C^t_j) - P(S_{i_k} C_j)}{P(S_{i_k} C_j)}
\tag{41}
$$

where by definition, $P(S_{i_k} C_j) = P(S_{i_k} C_j | C^t_j) P(C^t_j) + P(S_{i_k} C_j | C^f_j) P(C^f_j)$. Substituting in Equation (41), we get:

$$
\delta^t_{i_k j} = \frac{P(S_{i_k} C_j | C^t_j)(1 - P(C^t_j)) - P(S_{i_k} C_j | C^f_j) P(C^f_j)}{P(S_{i_k} C_j | C^t_j) P(C^t_j) + P(S_{i_k} C_j | C^f_j) P(C^f_j)}
\tag{42}
$$

Using the fact that $1 - P(C^t_j) = P(C^f_j)$ in the numerator, and rearranging, we get:

$$
\delta^t_{i_k j} = \frac{(P(S_{i_k} C_j | C^t_j) - P(S_{i_k} C_j | C^f_j)) P(C^f_j)}{P(S_{i_k} C_j | C^t_j) P(C^t_j) + P(S_{i_k} C_j | C^f_j) P(C^f_j)}
\tag{43}
$$

We can similarly show that:

$$
\begin{aligned}
\delta^f_{i_k j} &= \frac{P(S_{i_k} C_j | C^f_j) - P(S_{i_k} C_j)}{P(S_{i_k} C_j)} \\
&= \frac{P(S_{i_k} C_j | C^f_j)(1 - P(C^f_j)) - P(S_{i_k} C_j | C^t_j) P(C^t_j)}{P(S_{i_k} C_j | C^t_j) P(C^t_j) + P(S_{i_k} C_j | C^f_j) P(C^f_j)} \\
&= \frac{(P(S_{i_k} C_j | C^f_j) - P(S_{i_k} C_j | C^t_j)) P(C^t_j)}{P(S_{i_k} C_j | C^t_j) P(C^t_j) + P(S_{i_k} C_j | C^f_j) P(C^f_j)}
\end{aligned}
\tag{44}
$$

Dividing Equation (43) by Equation (44), we get:

$$
\frac{\delta^t_{i_k j}}{\delta^f_{i_k j}} = -\frac{P(C^f_j)}{P(C^t_j)}
\tag{45}
$$

Substituting for $\delta^t_{i_k j}$ from Equation (45) into Equation (40), we get $Terms_j = 0$. Substituting with this result in Equation (39), we get:

$$
P(Sources_j) = \prod_{k=1}^{K_j} P(S_{i_k} C_j)
\tag{46}
$$

The above result completes the proof. We have shown that the joint probability $P(S_{i_1} C_j, S_{i_2} C_j, ..., S_{i_K} C_j)$, denoted for simplicity by $P(Sources_j)$, is well approximated by the product of marginal probabilities $P(S_{i_1} C_j), ..., P(S_{i_K} C_j)$. Note that, the proof did not assume independence of the marginals. Instead, it proved the result under the small $\delta_{i_k j}$ assumption.

### REFERENCES

[1] J. Han, "Mining heterogeneous information networks by exploring the power of links," in *Proceedings of the 20th international conference on Algorithmic learning theory*, ser. ALT'09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 3–3. [Online]. Available: http://portal.acm.org/citation.cfm?id=1813231.1813235

[2] J. Pasternack and D. Roth, "Knowing What to Believe (when you already know something)," in *Proc. the International Conference on Computational Linguistics (COLING)*, 2010.

[3] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[4] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[5] X. Yin, P. S. Yu, and J. Han, "Truth Discovery with Multiple Conflicting Information Providers on the Web," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 6, pp. 796–808, 2008. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4415269

[6] A. Galland, S. Abiteboul, A. Marian, and P. Senellart, "Corroborating information from disagreeing views," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 131–140. [Online]. Available: http://portal.acm.org/citation.cfm?id=1718504

[7] M. Gupta, Y. Sun, and J. Han, "Trust analysis with clustering," in *WWW*, ser. WWW '11. New York, NY, USA: ACM, 2011.

[8] L. Berti-Equille, A. D. Sarma, X. Dong, A. Marian, and D. Srivastava, "Sailing the information ocean with awareness of currents: Discovery and application of source dependence," in *CIDR*, 2009.

[9] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Integrating conflicting data: The role of source dependence," *PVLDB*, vol. 2, no. 1, pp. 550–561, 2009.

[10] X. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava, "Global detection of complex copying relationships between sources," *PVLDB*, vol. 3, no. 1, pp. 1358–1369, 2010.

[11] L. Blanco, V. Crescenzi, P. Merialdo, and P. Papotti, "Probabilistic models to reconcile complex data from inaccurate data sources," in *CAiSE*, 2010, pp. 83–97.

[12] R. Balakrishnan, "Source rank: Relevance and trust assessment for deep web sources based on inter-source agreement," in *WWW*, ser. WWW '11. New York, NY, USA: ACM, 2011.

[13] X. Yin and W. Tan, "Semi-supervised truth discovery," in *WWW*. New York, NY, USA: ACM, 2011.