On Credibility Estimation Tradeoffs in Assured Social Sensing

Dong Wang, Lance Kaplan, Tarek Abdelzaher and Charu C. Aggarwal

***Abstract*—Two goals of network science are to (i) uncover fundamental properties of phenomena modeled as networks, and to (ii) explore novel use of networks as models for a diverse range of systems and phenomena in order to improve our understanding of such systems and phenomena. This paper advances the latter direction by casting *credibility estimation* in social sensing applications as a network science problem, and by presenting a network model that helps understand the fundamental accuracy trade-offs of a credibility estimator. Social sensing refers to data collection scenarios, where observations are collected from (possibly unvetted) *human* sources. We call such observations *claims* to emphasize that we do not know whether or not they are factually correct. Predictable, scalable and robust estimation of both source reliability and claim correctness, *given neither in advance*, becomes a key challenge given the unvetted nature of sources and lack of means to verify their claims. In a previous conference publication, we proposed a maximum likelihood approach to jointly estimate both source reliability and claim correctness. We also derived confidence bounds to quantify the accuracy of such estimation. In this paper, we cast credibility estimation as a network science problem and offer systematic sensitivity analysis of the optimal estimator to understand its fundamental accuracy trade-offs as a function of an underlying network topology that describes key problem space parameters. It enables *assured* social sensing, where not only source reliability and claim correctness are estimated, but also the accuracy of such estimates is correctly predicted for the problem at hand.**

***Index Terms*—Maximum Likelihood Estimation; Predictability; Cramer-Rao Lower Bound; Scalability; Robustness; Truth Discovery; Social Sensing;**

I. Introduction

THE REALIZATION that many distinct phenomena have underlying common network representations has recently spurred the emergence of network science as a discipline dedicated to the study of such networks and phenomena. Network science uncovers fundamental properties of networks and explores their use in modeling new, increasingly diverse natural and engineered systems and phenomena. This paper explores the problem of credibility estimation in social sensing applications, also known as *fact-finding* [11], [15], [18], [20], [21], as a network science problem. An underlying network representation is presented that describes the problem space. The impact of network topology is then studied on the fundamental performance trade-offs of the fact-finder.

Social sensing has emerged as an important sensing paradigm, where humans are explicitly or implicitly involved in data collection. Humans are generally less reliable than well-tested infrastructure sensors, and the correctness of their observations is often unknown *a priori*. Nevertheless, important decisions may need to be made based on collected data. To meet this challenge, a recent branch of machine learning literature (called *fact-finding*) [11], [15], [20], [21], addressed the problem of jointly estimating correctness of sources and claims, given neither in advance. This is in contrast to a large volume of past work where either source reliability was assumed to be known, or claims could be externally labeled as true or false by some training algorithm. In the absence of such knowledge, at the core of the new work is the idea of analyzing the topology of an *information network*, which is a graph whose nodes, at a minimum, represent sources and claims and whose edges denote who said what. Such an information network graph is depicted in Figure 1.

Understanding and quantifying the quality of information from topology analysis of the underlying information network is thus a network science problem. The analysis is cast as a problem of jointly estimating node attributes (namely, probability of correctness of each source and claim) given network topology. The general intuition behind this formulation lies in the observation that links between nodes act as constraints. Namely, they constrain the joint probability distribution of the attributes of nodes they connect. For example, the odds of correctness of a source and the odds of correctness of a claim it made are clearly related (and such relation constitutes the constraint). Hence, the topology of the information network yields the set of constraints that node attributes must obey. A solution to the fact-finding problem aims to find the assignment of node attributes (i.e., probability of correctness of sources and claims) that is maximally consistent with all constraints represented by the information network.

In preliminary conference publications [18]–[20], we developed an optimal solution to the fact-finding problem[1] based on a maximum likelihood estimation technique and analyzed its basic properties. We further quantified the confidence in estimation results based on the Cramer-Rao lower bound (CRLB) [19]. This paper completes the aforementioned work by exploring the impact of the underlying (information) network topology on fact-finder performance. Analytic expressions are used to study the sensitivity of estimator accuracy to changes in several aspects of topology of the underlying

Manuscript received August 16, 2012; revised January 31, 2013.

D. Wang and T. Abdelzaher are with the Department of Computer Science, University of Illinois, Urbana, IL 61801 (e-mail: {dwang24, zaher}@illinois.edu).

L. Kaplan is with the Networked Sensing & Fusion Branch, US Army Research Laboratory, Adelphi, MD 20783 (e-mail: lance.m.kaplan@us.army.mil).

C. C. Aggarwal is with IBM Research, Yorktown Heights, NY 10598 (e-mail: charu@us.ibm.com).

Digital Object Identifier 10.1109/JSAC.2013.130605.

[1]Under a simplifying set of assumptions as described in [20]

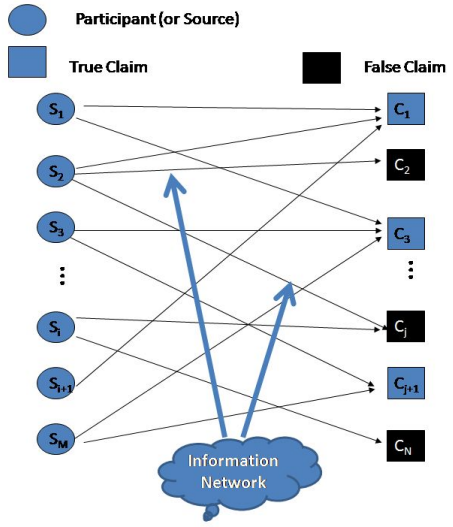


Fig. 1. The Fact-finding Information Network

information network, such as the number of nodes of different types, the number of edges, the distribution of edges, and the fraction of trusted nodes. We further validate the analytic results by both extensive simulation and a real-world social sensing application.

The results of this paper are important in two respects. First, while prior literature exists on information network analysis for purposes of fact-finding, these techniques do not offer an assessment of *quality of results*. In contrast, our approach not only provides the best hypothesis but also rigorously quantifies how good it is compared to ground truth. This quantification is immensely important in any practical settings, where errors have consequences. For example, in a military scenario, a response to the hypothesis that a particular organization harbors nuclear weapons can be vastly different depending on one's confidence in the hypothesis. *Estimating the confidence correctly and objectively from information network topology is therefore a key contribution to both fact-finding and network science.*

Second, sensitivity analysis of fact-finder accuracy (based on underlying network topology) is new to information network and data mining literature. Given our analytic quantification of confidence in results, we are able to rigorously analyze how such confidence changes as a function of information network topology (which reflects the input space of the fact-finding problem). Such sensitivity analysis offers a fundamental understanding of the capabilities and limitations of fact-finders.

The rest of this paper is organized as follows. In Section II, we briefly go over the maximum likelihood estimation (MLE) approach and the problem of quantifying result accuracy. We then derive actual and asymptotic bounds to compute confidence intervals in source reliability and estimate the number of misclassified claims (i.e., expected numbers of false positives and false negatives) in Section III. Evaluation results are presented in Section IV. We discuss the limitations of our model and possible extensions in Section V. Finally, we conclude the paper in Section VI.

II. Problem Statement

Our objective is explore the impact of information network topology on fact-finding accuracy in social sensing applications; a challenging problem due to the unknown reliability of data sources and the highly dynamic nature of social sensing topologies [1]. The basic fact-finders include Hubs and Authorities [11], Average.Log [15], and TruthFinder [21]. Other extended fact-finders and deception detection schemes further analyze properties of assertions and sources [2], [6]–[8], [22] and estimate the prevalence of deception or detect fraudsters in online communities [13], [14].

Recently, a Bayesian Interpretation scheme [17] was proposed to convert ranking outputs of fact-finders into probability of correctness. Wang *et al.* then proposed a maximum likelihood estimator, based on Expectation Maximization (EM) [20], that was shown to beat Bayesian Interpretation and other state-of-art fact-finders in estimation performance. A confidence interval, based on the Cramer-Rao lower bound (CRLB) [4], was computed for fact-finder output in a previous conference paper [19]. This lays the ground for an in-depth study of the impact of the underlying information network topology on optimal fact-finder performance.

Consider a social sensing application model, where a group of M sources, $S_1, ..., S_M$, make individual observations about a set of N claims $C_1, ..., C_N$ in their environment. For example, a group of local residents might join a geo-tagging campaign to report litter locations in a park. Hence, each claim denotes the existence or lack thereof of litter at a given location. We consider only binary claims and assume, without loss of generality, that their "normal" state is negative (e.g., no litter on the ground). Hence, sources report only when the positive state of the claim (e.g., litter found) is encountered. Each source generally observes only a small subset of all claims (e.g., states of places they have been to).

Let S_i denote the i^{th} source, C_j denote the j^{th} claim and S_iC_j denote that S_i reports C_j to be true. The social sensing topology describing *who reports what* can be represented by a bypartite *information network* SC, where source node S_i is connected to claim node C_j if S_i claims that C_j is true. Let $P(C_j^t)$ and $P(C_j^f)$ denote the odds that the actual claim C_j is indeed true and false, respectively. Let the probability that source S_i reports a claim be s_i (i.e., $s_i = P(S_iC_j)$ for all j). We refer to s_i as the *assertiveness* of the i^{th} source. Further, let the probability that source S_i is right be t_i. Note that, this probability represents the source's reliability, which is not known *a priori*. Formally, t_i is defined as:

$$t_i = P(C_j^t | S_iC_j) \tag{1}$$

Let a_i be the (unknown) probability that source S_i reports a claim, given that it is true, and b_i be the (unknown) probability that source S_i reports a claim, when it is actually false. Formally, a_i and b_i are defined as follows:

$$a_i = P(S_iC_j | C_j^t) \qquad b_i = P(S_iC_j | C_j^f) \tag{2}$$

Bayes' Theorem offers a relationship between t_i, a_i and b_i:

$$a_i = \frac{t_i \times s_i}{d} \qquad b_i = \frac{(1-t_i) \times s_i}{1-d} \tag{3}$$

where d is the overall prior probability that a randomly chosen claim is true.

Let us further define z_j such that it is 1 when claim C_j is true and 0 otherwise. A maximum-likelihood estimator can now take the information network SC as the input and iterate between the E-step and M-step of the EM scheme [5] until the estimation converges. An output of the EM scheme is the maximum likelihood estimation (MLE) of source reliability computed from its estimation parameter vector $\theta = (a_1, a_2, ...a_M; b_1, b_2, ...b_M)$. [2] Our goal is to: (i) derive the actual and asymptotic error bounds that characterize the accuracy of the maximum likelihood estimator and compute its confidence interval; (ii) estimate the accuracy of claim classification without knowing the ground truth values of the claims; and (iii) derive the dependency of the accuracy of maximum likelihood estimation on parameters of the problem space, as represented by the information network.

III. Performance analysis of the Maximum Likelihood Estimation

In this section, we analyze the accuracy of the maximum likelihood estimation in two ways: (i) we derive a confidence interval in source reliability estimates by computing the Cramer-Rao lower bounds (CRLBs) for the estimation parameters (i.e., θ) and leveraging the asymptotic normality of maximum likelihood estimation; (ii) we derive the expected number of misclassified claims (i.e., false claims classified as true and true claims classified as false). We further analyze the scalability of the actual CRLB derivation and suggest an asymptotic (approximate) CRLB that works for systems with a large number of sources.

A. Real Cramer Rao Lower Bound

We first derive the actual CRLB that characterizes the estimation accuracy of the maximum likelihood estimation of source reliability in social sensing. In estimation theory, the CRLB expresses a lower bound on the estimation variance of a minimum-variance unbiased estimator. In its simplest form, the bound states the variance of any unbiased estimator is at least as high as the inverse of the Fisher information [10]. The estimator that reaches this lower bound is said to be *efficient*. For notational convenience, we denote the information network SC as the observed data X and use $X_{ij} = S_i C_j$ for the following derivation.

The likelihood function (containing hidden variable Z) of the maximum likelihood estimation we get from EM can be expressed as [20]:

$$\begin{aligned} &L(\theta; X, Z) = p(X, Z|\theta) \\ &= \prod_{j=1}^{N} \left\{ \prod_{i=1}^{M} a_i^{X_{ij}} (1-a_i)^{(1-X_{ij})} \times d \times z_j \right. \\ &\left. + \prod_{i=1}^{M} b_i^{X_{ij}} (1-b_i)^{(1-X_{ij})} \times (1-d) \times (1-z_j) \right\} \end{aligned} \quad (4)$$

[2] In reality, the EM scheme can include the prior d in θ and jointly estimate its value [20]

where z_j is the hidden variable. The EM scheme is used to handle the hidden variable and aims to find:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(X|\theta) \quad (5)$$

where

$$\begin{aligned} p(X|\theta) = \prod_{j=1}^{N} \left\{ \prod_{i=1}^{M} a_i^{X_{ij}} (1-a_i)^{(1-X_{ij})} \times d \right. \\ \left. + \prod_{i=1}^{M} b_i^{X_{ij}} (1-b_i)^{(1-X_{ij})} \times (1-d) \right\} \end{aligned} \quad (6)$$

By definition of CRLB, it is given by

$$CRLB = J^{-1} \quad (7)$$

where

$$J = E[\bigtriangledown_\theta \ln p(X|\theta) \bigtriangledown_\theta^H \ln p(X|\theta)] \quad (8)$$

where J is the Fisher information of the estimation parameter, $\bigtriangledown_\theta = (\frac{\partial}{\partial a_1}, ... \frac{\partial}{\partial a_M}, \frac{\partial}{\partial b_1}, ..., \frac{\partial}{\partial b_M})^H$ and H denotes the conjugate transpose operation. In information theory, the Fisher information is a way of measuring the amount of information that an observable random variable X carries about an estimated parameter θ upon which the probability of X depends. The expectation in Equation (8) is taken over all values for X with respect to the probability function $p(X|\theta)$ for any given value of θ. Let $\mathcal{X}$ represent the set of all possible values of $X_{ij} \in \{0, 1\}$ for $i = 1, 2...M; j = 1, 2, ...N$. Note $|\mathcal{X}| = 2^{MN}$. Likewise, let $\mathcal{X}_j$ represent the set of all possible values of $X_{ij} \in \{0, 1\}$ for $i = 1, 2...M$ at a given value of j. Note $|\mathcal{X}_j| = 2^M$. Taking the expectation, Equation (8) can be rewritten as follows:

$$J = \sum_{X \in \mathcal{X}} \bigtriangledown_\theta \ln p(X|\theta) \bigtriangledown_\theta^H \ln p(X|\theta) p(X|\theta) \quad (9)$$

Then, the fisher information matrix can be represented as:

$$J = \begin{bmatrix} A & C \\ C^T & B \end{bmatrix}$$

where submatrices A, B and C contain the elements related with the estimation parameter a_i, b_i and their cross terms respectively. The representative elements A_{kl}, B_{kl} and C_{kl} of A, B and C can be derived as follows:

$$\begin{aligned} A_{kl} &= E\Big[\frac{\partial}{\partial a_k} \ln p(X|\theta) \frac{\partial}{\partial a_l} \ln p(X|\theta)\Big] \\ &= E\Big[\Big(\sum_j \frac{(2X_{kj}-1)Z_j}{a_k^{X_{kj}}(1-a_k)^{(1-X_{kj})}} \sum_q \frac{(2X_{lq}-1)Z_q}{a_l^{X_{lq}}(1-a_l)^{(1-X_{lq})}}\Big)\Big] \\ &= \sum_j \sum_q E\Big[\frac{(2X_{kj}-1)Z_j(2X_{lq}-1)Z_q}{a_k^{X_{kj}}(1-a_k)^{(1-X_{kj})} a_l^{X_{lq}}(1-a_l)^{(1-X_{lq})}}\Big] \end{aligned} \quad (10)$$

where

$$Z_j = p(z_j = 1|X) = \frac{A_j \times d}{A_j \times d + B_j \times (1-d)}$$

where

$$A_j = \prod_{i=1}^{M} a_i^{X_{ij}} (1-a_i)^{(1-X_{ij})} \quad B_j = \prod_{i=1}^{M} b_i^{X_{ij}} (1-b_i)^{(1-X_{ij})} \quad (11)$$

Z_j is the conditional probability of the claim C_j to be true given the information network, SC. After further simplification as shown in the appendix A, A_{kl} can be expressed as the sum of only the expectation terms where $j = q$:

$$A_{kl} = \sum_j E\Big[\frac{(2X_{kj}-1)(2X_{lj}-1)Z_j^2}{a_k^{X_{kj}}(1-a_k)^{(1-X_{kj})}a_l^{X_{lj}}(1-a_l)^{(1-X_{lj})}}\Big]$$
$$= \sum_{j=1}^{N}\sum_{X\in\mathcal{X}_j}\frac{(2X_{kj}-1)(2X_{lj}-1)\prod_{\substack{i=1\\i\neq k}}^{M}A_{ij}\prod_{\substack{i=1\\i\neq l}}^{M}A_{ij}d^2}{\prod_{i=1}^{M}A_{ij}d+\prod_{i=1}^{M}B_{ij}(1-d)} \tag{12}$$

where

$$A_{ij} = a_i^{X_{ij}}(1-a_i)^{(1-X_{ij})} \quad B_{ij} = b_i^{X_{ij}}(1-b_i)^{(1-X_{ij})} \tag{13}$$

Since the inner sum in (12) is invariant to the claim index j, we can rewrite $A_{k,l} = N\bar{A}_{k,l}$ where $\bar{A}_{kl}$ is:

$$\bar{A}_{kl} = \sum_{x\in\mathcal{X}_j}\frac{(2X_{kj}-1)(2X_{lj}-1)\prod_{\substack{i=1\\i\neq k}}^{M}A_{ij}\prod_{\substack{i=1\\i\neq l}}^{M}A_{ij}d^2}{\prod_{i=1}^{M}A_{ij}d+\prod_{i=1}^{M}B_{ij}(1-d)} \tag{14}$$

It should also be noted that the summation in Equation (14) is the same for all j.

By similar calculations, we can obtain the inverse of the Fisher information matrix as follows:

$$J^{-1} = \frac{1}{N}\begin{bmatrix}\bar{A} & \bar{C}\\ \bar{C}^T & \bar{B}\end{bmatrix}^{-1}$$

where we define the kl^{th} element of $\bar{B}$, $\bar{C}$ as:

$$\bar{B}_{kl} = \sum_{x\in\mathcal{X}_j}\frac{(2X_{kj}-1)(2X_{lj}-1)\prod_{\substack{i=1\\i\neq k}}^{M}B_{ij}\prod_{\substack{i=1\\i\neq l}}^{M}B_{ij}(1-d)^2}{\prod_{i=1}^{M}A_{ij}d+\prod_{i=1}^{M}B_{ij}(1-d)} \tag{15}$$

$$\bar{C}_{kl} = \sum_{x\in\mathcal{X}_j}\frac{(2X_{kj}-1)(2X_{lj}-1)\prod_{\substack{i=1\\i\neq k}}^{M}A_{ij}\prod_{\substack{i=1\\i\neq l}}^{M}B_{ij}d(1-d)}{\prod_{i=1}^{M}A_{ij}d+\prod_{i=1}^{M}B_{ij}(1-d)} \tag{16}$$

Note that the sum of $\bar{A}_{kl}$, $\bar{B}_{kl}$ and $\bar{C}_{kl}$ are over the 2^M different permutations of X_{ij} for $i = 1, 2, ...M$ at a given j. This is much smaller than the 2^{MN} permutations of $\mathcal{X}$.

This gives us the actual CRLB. Note that more claims simply lead to better estimates for θ as the variance decreases as $\frac{1}{N}$. The decrease in variance for the estimates as a function of M is more complicated. We can only compute it numerically.

B. Asymptotic Cramer Rao Lower Bound

Observe that the complexity of the actual CRLB computation in the above subsection is exponential with respect to the number of sources (i.e., M) in the system. Therefore, it is inefficient (or infeasible) to compute the actual CRLB when the number of sources becomes large. In this subsection, we outline the asymptotic CRLB for efficient computation in the sensing system with a large number of sources. The asymptotic CRLB is derived based on the assumption that the hidden variable (i.e., z_j) can be correctly estimated from EM, which is a reasonable assumption when the number of sources is sufficient. Under this assumption, the log-likelihood function of the maximum likelihood estimation we get from EM can be expressed as follows:

$$l_{em}(x;\theta) = \sum_{j=1}^{N}\Bigg\{ z_j \times \Bigg[\sum_{i=1}^{M}(X_{ij}\log a_i + (1-X_{ij})\log(1-a_i) + \log d)\Bigg] + (1-z_j) \times \Bigg[\sum_{i=1}^{M}(X_{ij}\log b_i + (1-X_{ij})\log(1-b_i) + \log(1-d))\Bigg]\Bigg\} \tag{17}$$

We first compute the Fisher Information Matrix at the MLE from the log-likelihood function given by Equation (17). According to prior work [20], the maximum likelihood estimator $\hat{\theta}_{MLE}$ is given by:

$$\hat{a}_i^{MLE} = \frac{\sum_{j=1}^{N}X_{ij}Z_j^c}{\sum_{j=1}^{N}Z_j^c} \quad \hat{b}_i^{MLE} = \frac{\sum_{j=1}^{N}X_{ij}(1-Z_j^c)}{N-\sum_{j=1}^{N}Z_j^c} \tag{18}$$

where Z_j^c is the converged probability of the j^{th} claim to be true from EM algorithm. Observe that each $\hat{a}_i^{MLE}$ or $\hat{b}_i^{MLE}$ is computed from N independent samples (i.e., claims).

Plugging $l_{em}(x;\theta)$ given by Equation (17) into the Fisher information defined in Equation (8), we have the representative element of Fisher Information Matrix from N claims as:

$$(J(\hat{\theta}_{MLE}))_{i,j} = \begin{cases}0 & i\neq j\\ -E_X\big[\frac{\partial^2 l_{em}(x;a_i)}{\partial a_i^2}|_{a_i=\hat{a}_i^{MLE}}\big] & i=j\in[1,M]\\ -E_X\big[\frac{\partial^2 l_{em}(x;b_i)}{\partial b_i^2}|_{b_i=\hat{b}_i^{MLE}}\big] & i=j\in(M,2M]\end{cases} \tag{19}$$

Substituting the log-likelihood function in Equation (17) and MLE in Equation (18) into Equation (19), the asymptotic CRLB (i.e., the inverse of the Fisher Information Matrix) can be written as:

$$(J^{-1}(\hat{\theta}_{MLE}))_{i,j} = \begin{cases}0 & i\neq j\\ \frac{\hat{a}_i^{MLE}\times(1-\hat{a}_i^{MLE})}{N\times d} & i=j\in[1,M]\\ \frac{\hat{b}_i^{MLE}\times(1-\hat{b}_i^{MLE})}{N\times(1-d)} & i=j\in(M,2M]\end{cases} \tag{20}$$

Note that the asymptotic CRLB is independent of M under the assumption that M is sufficient, and it can be quickly computed from the MLE of the EM scheme.

C. Confidence Interval

In this subsection, we show that the confidence interval of source reliability can be obtained by using the CRLB we derived in previous sections and leveraging the asymptotic normality of the maximum likelihood estimation.

The maximum likelihood estimator posses a number of attractive asymptotic properties. One of them is called *asymptotic normality*, which basically states the MLE estimator is

asymptotically distributed with Gaussian behavior as the data sample size goes up, in particular [3]:

$$(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N(0, J^{-1}(\hat{\theta}_{MLE})) \tag{21}$$

where J is the Fisher Information Matrix computed from all samples, θ_0 and $\hat{\theta}_{MLE}$ are the true value and the maximum likelihood estimation of the parameter θ respectively. The Fisher information at the MLE is used to estimate its true (but unknown) value [10]. Hence, the asymptotic normality property means that in a regular case of estimation and in the distribution limiting sense, the maximum likelihood estimator $\hat{\theta}_{MLE}$ is unbiased and its covariance reaches the Cramer-Rao lower bound (i.e., an efficient estimator).

From the asymptotic normality of the maximum likelihood estimator [4], the error of the corresponding estimation on θ follows a normal distribution with zero mean and the covariance matrix given by the CRLB we derived in previous subsections. Let us denote the variance of estimation error on parameter a_i as $var(\hat{a}_i^{MLE})$. Recall the relation between source reliability (i.e., t_i) and estimation parameter a_i and b_i is $t_i = \frac{a_i \times d}{a_i \times d + b_i \times (1-d)}$. For a sensing system with small values of M and N, the estimation of t_i has a complex distribution and its estimation variance can be approximated [4]. For a sensing system with sufficient M and N (i.e., under asymptotic condition), the denominator of t_i can be approximated as s_i based on Equation (3).[3] Therefore, $(\hat{t}_i^{MLE} - t_i^0)$ also follows a normal distribution with zero mean and variance given by:

$$var(\hat{t}_i^{MLE}) = \left(\frac{d}{s_i}\right)^2 var(\hat{a}_i^{MLE}) \tag{22}$$

Hence, we are now able to obtain the confidence interval that can be used to quantify the estimation accuracy of the maximum likelihood estimation on source reliability. The confidence interval of the reliability estimation of source S_i (i.e., $\hat{t}_i^{MLE}$) at confidence level p is given by the following:

$$(\hat{t}_i^{MLE} - c_p\sqrt{var(\hat{t}_i^{MLE})}, \hat{t}_i^{MLE} + c_p\sqrt{var(\hat{t}_i^{MLE})}) \tag{23}$$

where c_p is the standard score (z-score) of the confidence level p. For example, for the 95% confidence level, $c_p = 1.96$. Therefore, the derived confidence interval of the source reliability MLE, as we demonstrated, can be computed by using the CRLB derived in this section.

D. Estimation of Claim Classification Accuracy

In previous subsections, we discussed how to compute the CRLB and the confidence interval in source reliability from the maximum likelihood estimation (MLE) of the EM algorithm. However, one problem remains to be answered is how to estimate the accuracy of the claim classification (i.e, false positives and false negatives) without having the ground truth values of the claims at hand. In this subsection, we propose a quick and effective method to answer the above question under the maximum likelihood hypothesis.

The results of the EM algorithm not only offered the MLE on the estimation parameters (i.e., θ) but also the probability of each claim to be true given the observed data and estimation parameters, which is given by:

$$Z_j^* = p(z_j = 1 | X_j, \theta^*) \tag{24}$$

where X_j is the observed data of the claim C_j and θ^* is the maximum likelihood estimation of the parameter. Since the claim is binary, it is classified as true if $Z_j^* \geq 0.5$ and false otherwise. Based on the above definition, the false positives and false negatives in claim classification can be estimated as follows:

$$FP = \sum_{j:Z_j^* \geq 0.5}^{N} \{Z_j^* \times 0 + (1 - Z_j^*) \times 1\} = \sum_{j:Z_j^* \geq 0.5}^{N} (1 - Z_j^*) \tag{25}$$

$$FN = \sum_{j:Z_j^* < 0.5}^{N} \{Z_j^* \times 1 + (1 - Z_j^*) \times 0\} = \sum_{j:Z_j^* < 0.5}^{N} Z_j^* \tag{26}$$

where FP and FN stand for false positives and false negatives respectively. From the above equations, we can compute the estimated false positives and false negatives in claim classification under the maximum likelihood hypothesis. This enables us to estimate the accuracy of claim classification without knowing the ground truth values *a priori*.

In this section, we derived a confidence interval in source reliability and a claim classification accuracy estimator. This allows social sensing applications to assess the quality of their estimation results. In the following section, we evaluate the performance of the computed confidence bounds and estimated false positives and false negatives.

IV. Evaluation

In this section, we evaluate of the performance of our credibility estimation approach through both extensive simulation studies and a real world social sensing application. First, we built a simulator in Matlab 7.10.0 that generates a random number of sources and claims. A random probability P_i is assigned to each source S_i representing his/her reliability (i.e., the ground truth probability that they report correct observations). L_i claims are asserted by the i^{th} source, such that the probability of true claims is P_i. We let P_i be uniformly distributed between 0.5 and 1 in our experiments[4]. The prior, d, discussed in Section II is set to 0.5 unless otherwise specificed and the initial value of d assumed by the EM alogrithm is uniformly distributed between 0.4 and 0.6.

A. Evaluation of Confidence Intervals

In this subsection, we evaluate the accuracy of computing the confidence interval in source reliability. We carried out experiments over three different information network scales: small, medium and large, featuring 100, 1000, and 10000 sources, respectively. In each case, half of the claimed items were true. We ran the EM algorithm and computed the confidence interval in reliability of each individual source

[3]The value of s_i can be estimated as $\frac{L_i}{N}$, where L_i is the number of observations reported by source S_i

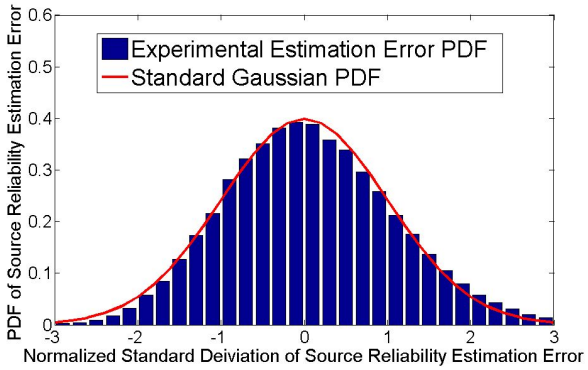


Fig. 2. Normalized Source Reliability Estimation Error PDF

TABLE I
ACCURACY OF CONFIDENCE INTERVALS

	68%	90%	95%
100 nodes	67.11%	89.13%	94.37%
1000 nodes	67.85%	89.49%	94.51%
10000 nodes	68.35%	89.85%	94.82%

based on Equation (23). We repeated the experiments 100 times for statistical significance.

Figure 2 shows the normalized probability density function (PDF) of source reliability estimation error for a network of 1000 sources. (Results for 100 and 10000 source networks are similar and not shown.) Comparing the experimental PDF to the standard Gaussian distribution, we verify that the asymptotic normality property holds.

Table I demonstrates the accuracy of computed confidence intervals on source reliability. It shows (for each given confidence interval) the percentage of sources whose reliability actually stays within the interval. Note that, the latter is indeed approximately equal to the former, which demonstrates the accuracy of interval estimation. The comparison is made for networks of size 100, 1000, and 10000, and for confidence levels of 68%, 90%, and 95%. The table shows the average of 100 experiments per cell.

B. Evaluation of CRLB

In this subsection, we evaluate the accuracy of our CRLBs (both the actual and asymptotic) at bounding estimation parameter error, as derived in Section III-A and III-B. Specifically, we compare these bounds to the actual variance of the corresponding estimation parameters. We focus on parameter a_i (the probability that source S_i reports a claim, given that it is true). Results for parameter b_i are similar. The actual estimation variance is characterized by the average RMSE (root mean square estimation error) over all sources.

1) Scalability Study: We first evaluate the accuracy of CRLBs with respect to network size (i.e, M and N). The first experiment evaluates the effect of the number of sources (i.e., M) in the network on the accuracy of both the actual and asymptotic CRLBs. Reported results are averaged over 100 experiments and are shown in Figure 3. Observe that the actual

[4]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

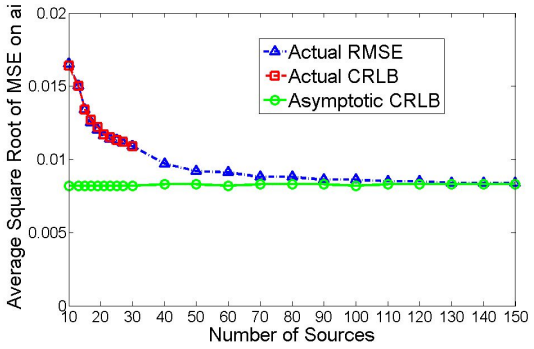


Fig. 3. CRLB versus Varying M

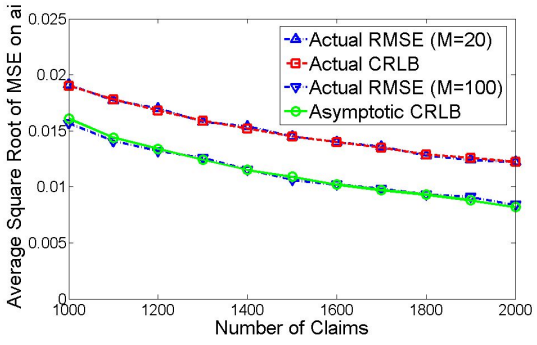


Fig. 4. CRLB versus Varying N

CRLB tracks the variance of estimation parameters accurately even when the number of sources is small (e.g., $M \leq 20$) in the system. We also observe that the RMSE is smaller than the actual CRLB when there are too few sources. This is because the estimator is biased for small datasets. The asymptotic CRLB deviates more from the actual estimation variance when the number of sources is small (e.g., $M \leq 20$). However, as the number of sources becomes sufficient in the network, the RMSE converges to the asymptotic CRLB quickly and the difference between the two becomes insignificant.

The second experiment compares the derived CRLBs (both actual and asymptotic) to the RMSE of estimation parameters when the number of claims (i.e., N) changes. As shown in Section III, both the actual and asymptotic CRLB decrease as $\frac{1}{N}$. As before, we observe that the actual CRLB is able to track the RMSE on estimation parameter correctly and they both decrease approximately as $\frac{1}{N}$ when the number of claim increases. Similarly, we observe that the asymptotic CRLB follows closely the RMSE of the estimation parameter when the number of claims increases.

2) Trustworthiness and Assertiveness Study: Next, we evaluate the accuracy of CRLB when the ratio of trusted sources in the system changes. The trusted sources are the sources who always make correct observations (i.e, their reliability is 1). We vary the trusted source ratio from 0 to 0.9. Reported results are averaged over 100 experiments and shown in Figure 5. Observe that both the actual and asymptotic CRLBs track the estimation variance tightly. We also note that both the CRLBs and the variance of estimation parameters improve as the trusted source ratio increases, as might be expected.

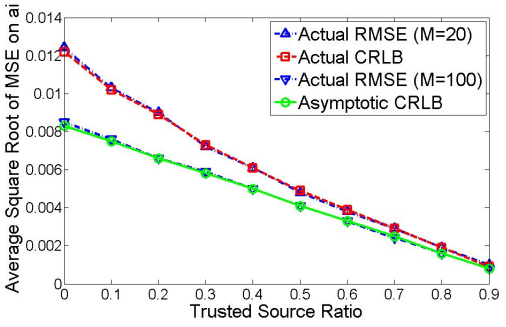


Fig. 5. CRLB versus Trusted Source Ratio

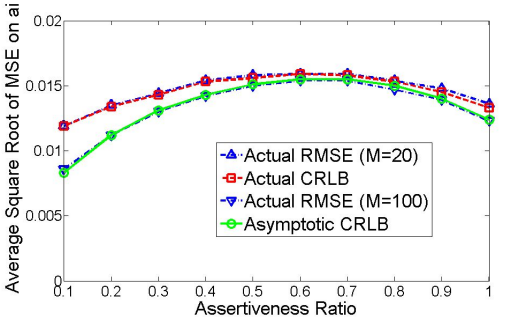
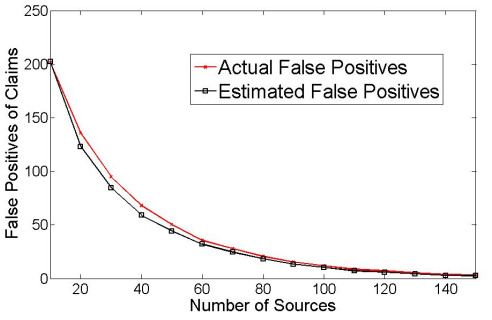
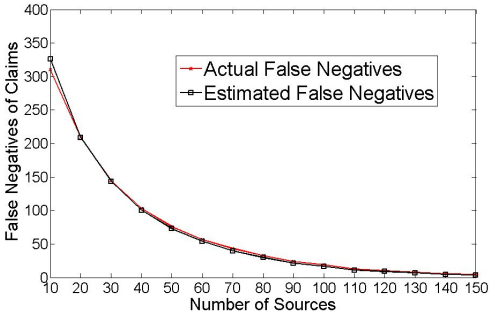


Fig. 6. CRLB versus Assertiveness Ratio



(a) False Positives



(b) False Negatives

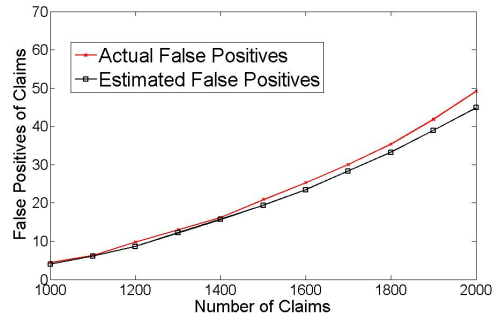
Fig. 7. Estimation of Claim Classification Accuracy versus Varying M

Finally, in the assertiveness study, we evaluate the accuracy of CRLBs when the assertiveness ratio of sources changes. The assertiveness ratio of a source is a measure of the number of observations it makes. An assertiveness ratio of 1 corresponds to 1000 observations per source over the duration of the experiment. We vary the assertiveness ratio from 0.1 to 1. Reported results are averaged over 100 experiments and shown in Figure 6. We observe that both the actual and asymptotic CRLBs track the RMSE of the estimation parameters correctly as the assertiveness ratio changes.

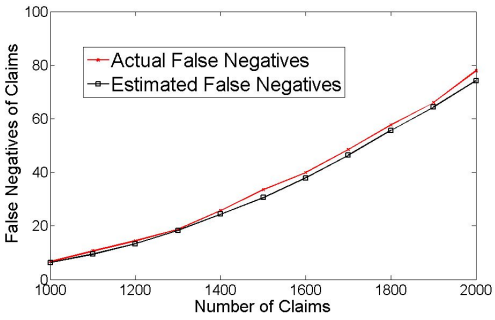
C. Evaluation of Estimated False Positives/Negatives on Claim Classification

In this subsection, we evaluate the estimated false positives/negatives on claim classification derived in Section III-D by comparing them to the actual false positives/negatives (i.e, the ones that are computed from the ground truth). We carried out similar experiments to the previous subsection and evaluated scalability, trustworthiness, and assertiveness.

1) Scalability Study: We first evaluate the scalability of the estimated false positives/negatives with respect to the sensing topology. The first experiment evaluates the performance when the number of sources (i.e., M) in the system changes. We fix the number of true and false claims at 1000. The average number of observations per source is set to 200. We vary the number of sources from 10 to 150. Reported results are averaged over 100 experiments and are shown in Figure 7. Observe that both estimated false positives and false negatives track the actual values accurately as the number of sources changes. We also note that the false positives/negatives decrease as the number of sources increases. The second experiment compares the estimated false positives/negatives to



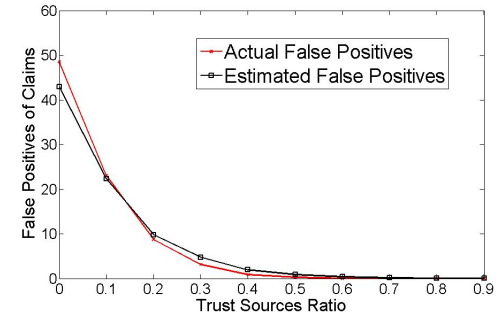
(a) False Positives



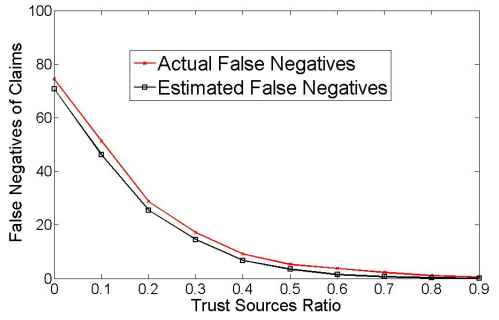
(b) False Negatives

Fig. 8. Estimation of Claim Classification Accuracy versus Varying N

the actual values when the number of claims (i.e., N) changes. We fix the number of sources at 50. The average number of observations per source is set to 200. We also keep the number of true and false claims the same. We vary the number of claims from 1000 to 2000. Reported results are averaged over 100 experiments and shown in Figure 8. Observe that

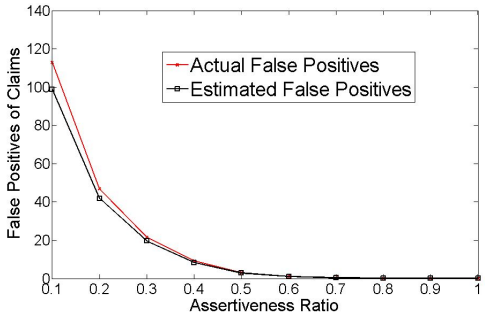


(a) False Positives

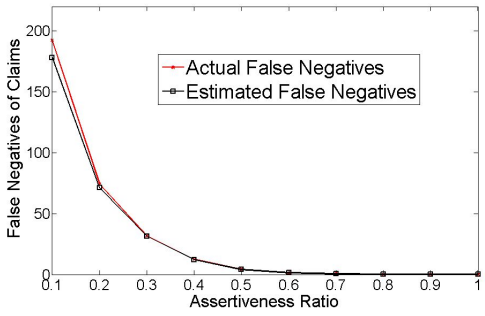


(b) False Negatives

Fig. 9. Estimation of Claim Classification Accuracy versus Trusted Sources Ratio



(a) False Positives



(b) False Negatives

Fig. 10. Estimation of Claim Classification Accuracy versus Assertiveness Ratio

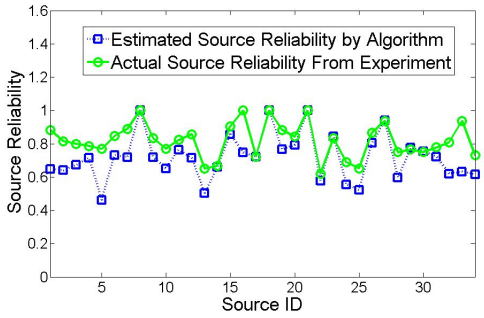
the estimated false positives/negatives are able to track the actual values correctly when the number of claims changes. We also note that the estimation performance degrades as the number of claims increases. The reason is: the sensing topology becomes sparser as the number of claims increases while the number of sources and observations per source stay the same.

2) Trustworthiness and Assertiveness Study: In the trustworthiness study, we evaluate the estimated false positives/negatives when the ratio of trusted sources changes in the system. In the experiment, we fix the number of sources to be 50. The number of true and false claims are set to 1000. The observations per source are set to 200. We vary the trusted source ratio from 0 to 0.9. The reported results are averaged over 100 experiments and shown in Figure 9. Observe that the estimated false positives/negatives track the actual values correctly and both of them decrease as the trusted source ratio increases. The reason is: trusted sources always provide correct observations, which helps the algorithm estimate the truthfulness of claims more accurately.

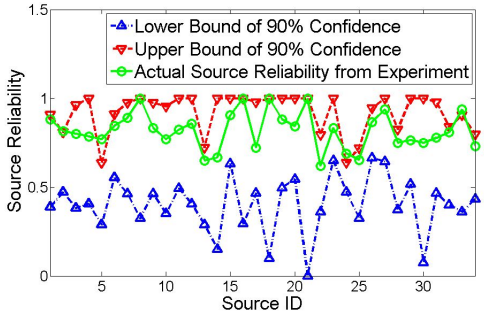
In the assertiveness study, we evaluate the estimated false positives/negatives when the assertiveness ratio changes in the system. In the experiment, we fix the number of sources at 50. The number of true and false claims are set to 1000. We vary the assertiveness ratio from 0.1 to 1. The reported results are averaged over 100 experiments and shown in Figure 10. Observe that the estimated false positives/negatives track the actual values correctly and both of them decrease as the assertiveness ratio increases. The reason is: the sensing topology becomes more densely connected and offers a better chance for the algorithm to correctly judge the truthfulness of the claims as the assertiveness ratio increases.

D. A Real World Case Study

In this section, we evaluate the performance of our credibility estimation approach through a real-world social sensing application. The goal of this application is to identify the correct locations of traffic lights and stop signs in the twin city of Urbana-Champaign by leveraging GPS devices on a set of vehicles traveling regularly in town. (The identified traffic light and stop sign locations were then used along with other information to compute fuel and delay estimates on city routes for a recent green navigation service [9]). We distributed Google's Galaxy Nexus Android phones to a group of participants who agreed to put them in their cars. Our test application, on the phone, recorded GPS traces, where every GPS reading is composed of an instantaneous latitude-longitude location, speed, time, and bearing of the vehicle. The application then computed simple features that constitute (intentionally unreliable) indicators that the vehicle is waiting at stop sign or a traffic light. Specifically, if a vehicle stops at a location for 15-90 seconds, the application concludes that it is stopped at a traffic light at that location. Similarly if it stops for 2-10 seconds, it concludes that it is at a stop sign. These conclusions were reported as claims from the corresponding source. The claim would be that a stop sign (or a traffic light, as applicable) exists at the current location and bearing. Clearly, these generated claims are unreliable, due to the simple-minded nature of the "sensor" and the complexity of road conditions and driver's behaviors. For example, a car can stop anywhere on the road due to a traffic jam or a crossing pedestrian, not necessarily at the location of traffic lights or stop signs. Also, cars do not stop at traffic lights when they are green. Finally, different drivers have different attitudes towards

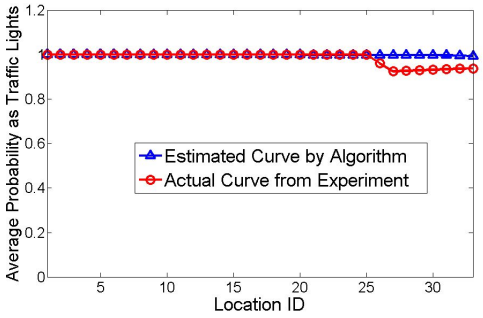


(a) Source Reliability Prediction

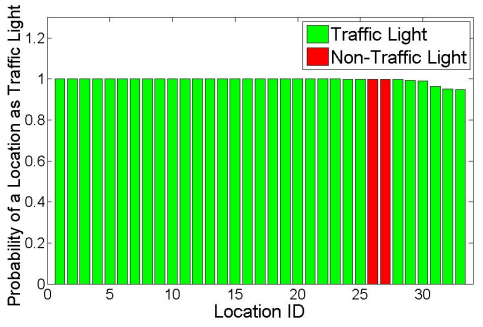


(b) Source Reliability Bound

Fig. 11. Estimation of Source Reliability in the Case of Traffic Lights



(a) Claim Correctness Prediction



(b) Claim Classification

Fig. 12. Estimation of Claim Correctness in the Case of Traffic Lights

stop signs. Some are more careless and may pass stop signs without stopping or do a "rolling stop", whereas others reliably stop at each sign.

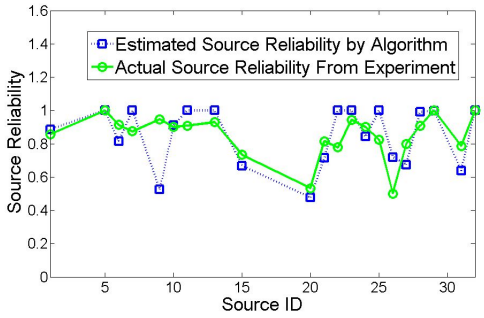
The general lack of reliability of claims and sources (and the differences in driver behavior) constituted a good test for the fact-finding algorithm described in this paper. Hence, we applied our credibility estimation approach to the collected claim data with the hope to find the correct locations of traffic lights and stop signs, and to identify the reliability of participants. For evaluation purposes, we also independently manually collected ground truth locations of traffic lights and stop signs.

In the experiment, 34 people were invited to participate in the application and 1,048,572 GPS readings (around 300 hours of driving) were collected. A total of 4865 claims were generated by the phones, of which 3303 were for stop signs and 1562 were for traffic lights, collectively identifying 369 distinct locations. We then generated the information network by taking the participants as sources and their stop sign/traffic light reports as claims. We applied the proposed credibility estimation approach to the data collected and evaluated its accuracy at inferring which reports were correct.

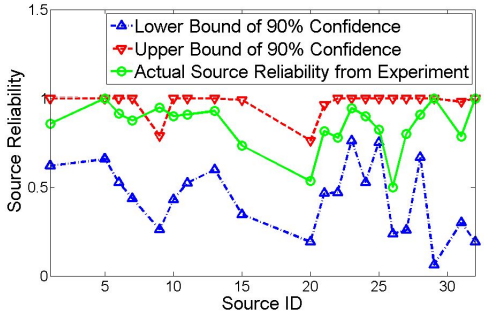
Figure 11 and Figure 12 show the results for the case of traffic lights identification. Figure 11 shows the results of source reliability. In Figure 11(a), we compared the source reliability estimated by our credibility estimation algorithm with the actual source reliability (i.e., the percentage of claims from that source that were actually correct) computed from ground truth. We observed that estimated values track actual results well for most of the sources. Figure 11(b) shows the 90% confidence bounds on source reliability estimation. We observe that actual source reliablity stays moslty within the 90% confidence bound. Only 3 sources out of 34 (less than 10%) have their reliability slightly outside the bound, which is what a 90% confidence interval means. Hence, the experiment verifies the accuracy and tightness of the confidence bounds derived in Section III. We also examined the 68% and 95% confidence bounds and observed that they caputre the 70.6% and 94.1% of sources whose relability estimations stay within bounds, respectively. This again verifies accuracy of those confidence intervals. Figure 12 shows the results of claim classficiation on traffic lights. We sorted all locations, where the system identified traffic lights (i.e., concluded that the corresponding claims were true), by the probability of correctness, also returned by the system. We expect that traffic light locations identified with a higher probability will tend to be real lights, whereas those identified with a lower probability will include progressively more false positives.

Figure 12(a) shows the sorted locations on the x-axis, and computes for each n, the average probability that the first n locations are traffic lights. We compare the estimated probability to the actual ground truth probability. We observe that our estimation follows quite well the actual experimental ground truth. It verifies the accuracy of the probability values computed and used for claim classification. Additionally, Figure 12(b) shows for each traffic light, in the same sorted order, the actual location status (i.e., whether a traffic light is in fact present at the location or not). We observe that most of the traffic light locations identified by our scheme are correct, although false positives arise as we go down in location ranking.

We repeated the above experiments for stop sign identification and observed similar trends as we had for traffic lights. However, we do find the identification of stop signs

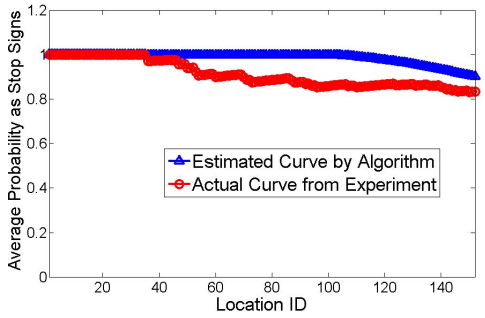


(a) Source Reliability Prediction

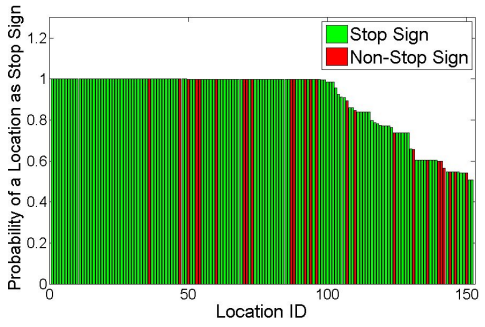


(b) Source Reliability Bound

Fig. 13. Estimation of Source Reliability in the Case of Stop Signs



(a) Claim Correctness Prediction



(b) Claim Classification

Fig. 14. Estimation of Claim Correctness in the Case of Stop Signs

more challenging than that of traffic lights. The reasons are: (i) the corroborated data for stop signs is sparser because the chances of different cars to stop at the same stop sign are much lower than that for traffic lights; (ii) cars have quite a few short wait behaviors at non-stop sign locations such as exists from parking lots, left turns, and pedestrian crossings; (iii) cars' bearings are usually not well aligned with the directions of stop signs, which is especially true when the car wants to make a turn after the stop sign. Therefore, for the evaluation of stop signs, we only picked sources whose reliability was more than 50%. Figure 13 shows the estimation results of source reliability. We observe that the actual source reliablity is estimated accurately and bounded correctly by the 90% confidence bounds. Figure 14 shows the results of claim classification at stop signs. We observe that the actual probability of correctness curve stays close to but slightly lower than the estimated one. The reasons of such deviation can be explained by the short wait behaviors mentioned above at non-stop sign locations in real world scenarios. In a sense, given our wait-based features, our algorithm actually did a better job at identifying actual *stop* locations of vehicles than would be predicted by looking at stop signs only. For example, it also found exits from parking lots and locations of pedestrian corssings. Note that, the aforementioned false positives gradually appear at locations that are ranked lower by the algorithm.

V. Limitations and Future Work

This paper analyzes the accuracy of maximum likelihood estimation in social sensing. Several simplifying assumptions were made that offer opportunities for future work.

Sources were assumed to be independent. In reality, sources could be influenced by each other. For example, they may copy observations, forward rumors, or even collude to misrepresent the truth. Hence, fact-finding should be cognizant of the social network among the information sources, as such network offers pathways for information propagation that violates the independent sources assumption made in our maximum likelihood estimation. The interaction between the social and information networks in social sensing is shown in Figure 15.

Recent work has proposed techniques to detect the dependency and copying relationship between sources [7]. Other methods are proposed to mitigate the source collusion attack by analyzing the network or interaction pattern of colluding sources [12]. In sociology, Exponential Random Graph Model (ERGM) has been widely used to study the interdependence of sources in social networks. ERGM can represent structural tendencies and define complicated source dependence patterns that are not easily captured by basic probabilistic models [16]. The above techniques can be used together with our quantification scheme to handle source dependency. The authors are currently working on extending the current model to handle non-independent sources. For example, one could cluster dependent sources into approximately independent clusters according to some source similarity metric [2] and run our scheme on top of the clustered sources. Additionally, sources are sometimes experts in specific domains. It would be interesting to assess estimation performance when taking source expertise into consideration. One possibility is to weight observations differently depending on the source's expertise.

No dependencies were assumed among different claims. There may be cases, however, observations on one claim could

imply observations on another (e.g., "flooding" at city B may imply "raining" at city A). Knowledge of such dependencies can thus be integrated with our scheme to pre-process the information network based on the reported observations and their relationships. Moreover, all observations are treated equally in our model. It is interesting to extend the model to handle the hardness of different observations. In other words, source reliability and confidence estimation may be computed not only based on whether those observations from the source are true or not but also based on whether such observations are trivial to make. This extension prevents sources from obtaining a track record of high reliability by making many trivially true observations. There are techniques that analyze the hardness of observations, which may be integrated with our scheme [8]. In this paper, sources are assumed to report only positive states of binary claims (e.g., litter found). This is a reasonable assumption for many social sensing applications (e.g., geotagging) where states of the observed variables is either true or false. However, sources can also make contradicting observations and claims can be non-binary in other types of applications (e.g., an on-line review system). Our model can be extended to handle contradicting observations as well as non-binary claims by expanding the estimation parameter vector that covers only positive states to every possible state of the claim and rebuilding the likelihood function. The authors are currently working on the above extensions.

VI. Conclusion

This paper studied the fundamental accuracy trade-offs in source and claim credibility estimation in social sensing applications. Our results allow applications to not only assess the reliability of sources and claims, given neither in advance, but also estimate the accuracy of such assessment. Confidence bounds on source reliability are computed based on the Cramer-Rao lower bound (CRLB). The accuracy of claim classification is estimated by computing the probability that each claim is correct. The derived accuracy results are shown to predict actual errors very well. The paper is a step towards assured social sensing: sensing that relies on unvetted sources, where nevertheless guarantees can be given on accuracy of fact-finding conclusions. We also note that our problem formulation and the proposed techniques are general enough to apply to any bi-partite network abstraction. The authors are currently making efforts to generalize the technique to solve problems beyond fact-finding.

Acknowledgements

Research reported in this paper was sponsored by the Army Research Laboratory, DTRA grant HDTRA1-10-1-0120, and NSF grants CNS 1040380 and CNS 09-05014, and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

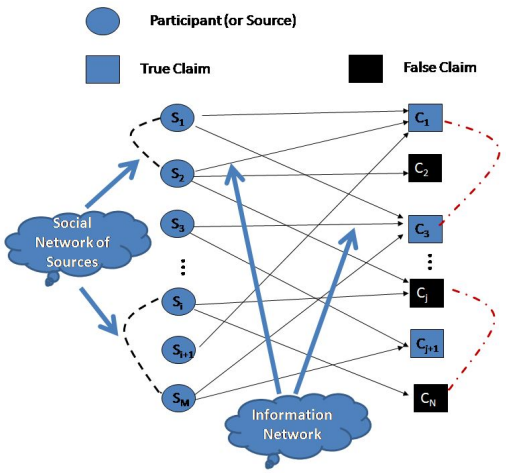


Fig. 15. Extended Social Sensing Information Network

Appendix A

When $j \neq q$, plugging the expressions of Z_j and Z_q, we can prove the expectation term in Equation (10) is zero:

$$
\begin{aligned}
&E\Big[\frac{(2X_{kj}-1)Z_j(2X_{lq}-1)Z_q}{a_k^{X_{kj}}(1-a_k)^{(1-X_{kj})}a_l^{X_{lq}}(1-a_l)^{(1-X_{lq})}}\Big] = \\
&\sum_{x\in\mathcal{X}}(2X_{kj}-1)(2X_{lq}-1)\times \\
&\Big(\prod_{\substack{i=1\\ i\neq k}}^{M} a_i^{X_{ij}}(1-a_i)^{(1-X_{ij})}\times d\prod_{\substack{i=1\\ i\neq l}}^{M} a_i^{X_{iq}}(1-a_i)^{(1-X_{iq})}\times d\Big) \\
&\times\Bigg(\prod_{\substack{j'=1\\ j'\neq j \text{ or } q}}^{N}\Big\{\prod_{i=1}^{M} a_i^{X_{ij'}}(1-a_i)^{(1-X_{ij'})}\times d \\
&+\prod_{i=1}^{M} b_i^{X_{ij'}}(1-b_i)^{(1-X_{ij'})}\times(1-d)\Big\}\Bigg) \\
&=\sum_{x\in\mathcal{X}_j\times\mathcal{X}_q}(2X_{kj}-1)(2X_{lq}-1)\times \\
&\Big(\prod_{\substack{i=1\\ i\neq k}}^{M} a_i^{X_{ij}}(1-a_i)^{(1-X_{ij})}\times d\prod_{\substack{i=1\\ i\neq l}}^{M} a_i^{X_{iq}}(1-a_i)^{(1-X_{iq})}\times d\Big) \\
&=\sum_{X_{kj}=0}^{1}\sum_{X_{lq}=0}^{1}(2X_{kj}-1)(2X_{lq}-1)=0 \qquad j\neq q
\end{aligned}
\tag{27}
$$

References

[1] C. Aggarwal and T. Abdelzaher. Integrating sensors and social networks. *Social Network Data Analytics, Springer*, 2011.
[2] 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'09*, 2009.
[3] G. Casella and R. Berger. *Statistical Inference*. Duxbury Press, 2002.
[4] H. Cramer. *Mathematical Methods of Statistics*. Princeton Univ. Press., 1946.
[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.
[6] X. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava. Global detection of complex copying relationships between sources. *PVLDB*, 3(1):1358–1369, 2010.
[7] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. *Proc. VLDB Endow.*, 2:550–561, August 2009.

[8] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *WSDM*, pages 131–140, 2010.
[9] R. K. Ganti, N. Pham, H. Ahmadi, S. Nangia, and T. F. Abdelzaher. Greengps: a participatory sensing fuel-efficient maps application. In *MobiSys '10: Proceedings of the 8th international conference on Mobile systems, applications, and services*, pages 151–164, New York, NY, USA, 2010. ACM.
[10] R. V. Hogg and A. T. Craig. *Introduction to mathematical statistics*. Prentice Hall, 1995.
[11] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.
[12] Q. Lian, Z. Zhang, M. Yang, B. Y. Zhao, Y. Dai, and X. Li. An empirical study of collusion behavior in the maze p2p file-sharing system. In *Proc. 27th International Conference on Distributed Computing Systems (ICDCS '07)*, 2007.
[13] M. Ott, C. Cardie, and J. Hancock. Estimating the prevalence of deception in online review communities. In *Proc. 21st international conference on World Wide Web*, WWW '12, pages 201–210, New York, NY, USA, 2012. ACM.
[14] S. Pandit, D. H. Chau, S. Wang, and C. Faloutsos. Netprobe: a fast and scalable system for fraud detection in online auction networks. In *Proc. 16th international conference on World Wide Web*, WWW '07, pages 201–210, New York, NY, USA, 2007. ACM.
[15] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *International Conference on Computational Linguistics (COLING)*, 2010.
[16] T. A. B. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock. New specifications for exponential random graph models. *Sociological Methodology*, 36(1):99–153, 2006.
[17] D. Wang, T. Abdelzaher, H. Ahmadi, J. Pasternack, D. Roth, M. Gupta, J. Han, O. Fatemieh, and H. Le. On bayesian interpretation of fact-finding in information networks. In *14th International Conference on Information Fusion (Fusion 2011)*, 2011.
[18] D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal. Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications. In *The 33rd International Conference on Distributed Computing Systems (ICDCS'13)*, July 2013.
[19] D. Wang, L. Kaplan, T. Abdelzaher, and C. C. Aggarwal. On scalability and robustness limitations of real and asymptotic confidence bounds in social sensing. In *The 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON 12)*, June 2012.
[20] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher. On truth discovery in social sensing: A maximum likelihood estimation approach. In *The 11th ACM/IEEE Conference on Information Processing in Sensor Networks (IPSN 12)*, April 2012.
[21] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE Trans. Knowl. Data Eng.*, 20:796–808, June 2008.
[22] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *Proc. VLDB Endow.*, 5(6):550–561, Feb. 2012.

Dong Wang received his Ph.D degree in Computer Science from University of Illinois at Urbana Champaign (UIUC) in December 2012 and currently work there as a postdoctoral researcher. Dong got his Master's degree in Electrical and Computer Engineering from Peking University (PKU), and BEng degree in Communication and Information Systems from University of Electronic Science and Technology of China (UESTC). Dr Wang's primary research focus is to analyze and understand the Quality of Information (QoI) in the emerging area of Social (Human-Centric) Sensing, which lies at the intersection between Cyber-Physical Systems and Information Networks. Dr. Wang is also interested in optimizing the tradeoffs between the QoI and data collection costs in data fusion system and designing energy efficient scheduling scheme for heterogeneous platform in sensor network. Dr Wang held the Wing Kai Cheng fellowship from University of Illinois. He is the recipient of the best paper award of Real-Time and Embedded Technology and Applications Symposium (RTAS) 2010.

Lance M. Kaplan received the B.S. degree with distinction from Duke University, Durham, NC, in 1989 and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1991 and 1994, respectively, all in Electrical Engineering. From 1987-1990, Dr. Kaplan worked as a Technical Assistant at the Georgia Tech Research Institute. He held a National Science Foundation Graduate Fellowship and a USC Dean's Merit Fellowship from 1990-1993, and worked as a Research Assistant in the Signal and Image Processing Institute at the University of Southern California from 1993-1994. Then, he worked on staff in the Reconnaissance Systems Department of the Hughes Aircraft Company from 1994-1996. From 1996-2004, he was a member of the faculty in the Department of Engineering and a senior investigator in the Center of Theoretical Studies of Physical Systems (CTSPS) at Clark Atlanta University (CAU), Atlanta, GA. Currently, he is a researcher in the Networked Sensing and Fusion branch of the U.S. Army Research Laboratory. Dr. Kaplan serves as Editor-In-Chief for the IEEE Transactions on Aerospace and Electronic Systems (AES). In addition, he also serves on the Board of Governors of the IEEE AES Society and on the Board of Directors of the International Society of Information Fusion. He is a three time recipient of the Clark Atlanta University Electrical Engineering Instructional Excellence Award from 1999-2001. His current research interests include signal and image processing, automatic target recognition, information/data fusion, and resource management.

Tarek Abdelzaher received his Ph.D. from the University of Michigan, Ann Arbor, in 1999, under Professor Kang Shin. He was an Asistant Professor at the University of Virginia from August 1999 to August 2005. He then joined the University of Illinois at Urbana Champaign as an Associate Professor with tenure, where he became Full Professor in 2011. His interests lie primarily in systems, including operating systems, networking, sensor networks, distributed systems, and embedded real-time systems. Dr Abdelzaher is especially interested in developing theory, architectural support, and computing abstractions for predictability in software systems, motivated by the increasing software complexity and the growing sources of non-determinism. Applications range from sensor networks to large-scale server farms, and from transportation systems to medicine.

Charu Aggarwal is a Research Scientist at the IBM T. J. Watson Research Center in Yorktown Heights, New York. He completed his B.S. from IIT Kanpur in 1993 and his Ph.D. from Massachusetts Institute of Technology in 1996. His research interest during his Ph.D. years was in combinatorial optimization (network flow algorithms), and his thesis advisor was Professor James B. Orlin . He has since worked in the field of performance analysis, databases, and data mining. He has published over 200 papers in refereed conferences and journals, and has applied for or been granted over 80 patents. He has received several invention achievement awards and has thrice been designated a Master Inventor at IBM. He is a recipient of an IBM Corporate Award (2003) for his work on bio-terrorist threat detection in data streams, a recipient of the IBM Outstanding Innovation Award (2008) for his scientific contributions to privacy technology, and a recipient of an IBM Research Division Award (2008) for his scientific contributions to data stream research. He served as an associate editor of the IEEE Transactions on Knowledge and Data Engineering Journal from 2004 to 2008. He is an associate editor of the ACM TKDD Journal , an action editor of the Data Mining and Knowledge Discovery Journal , an associate editor of the ACM SIGKDD Explorations, and an associate editor of the Knowledge and Information Systems Journal. He is a fellow of the IEEE for "contributions to knowledge discovery and data mining techniques", and a life-member of the ACM.