On Scalability and Robustness Limitations of Real and Asymptotic Confidence Bounds in Social Sensing

Dong Wang*, Lance Kaplan‡, Tarek Abdelzaher*† and Charu C. Aggarwal§
*Department of Computer Science, University of Illinois, Urbana, IL 61801
Email: dwang24, zaher@illinois.edu
†Department of Automatic Control, Lund University, Lund, Sweden (Sabbatical Affiliation)
‡Networked Sensing & Fusion Branch, US Army Research Laboratory, Adelphi, MD 20783
Email: lance.m.kaplan@us.army.mil
§IBM Research, Yorktown Heights, NY 10598
Email: charu@us.ibm.com

***Abstract*—This paper estimates new confidence bounds on source reliability in social sensing applications. Scalable and robust estimation of source reliability is a key challenge in social sensing where humans or human-operated sensors act as data sources. In order to assess correctness of data, the reliability of sources must first be assessed, yet this is complicated when sources are not a priori known and vetted, but rather can opt in at will, for example, by downloading a sensing application on their mobile device. In our previous work, we developed a maximum likelihood source reliability estimator and approximately quantified confidence in its estimation based on an asymptotic Cramer-Rao lower bound (CRLB). In this paper we show that the asymptotic bound fails to track estimation performance when the number of sources is small. We derive the real CRLB to accurately characterize estimation performance for scenarios where the asymptotic bound fails. We study the limitations of the real and asymptotic CRLBs and show the trade-offs they offer between computational complexity and estimation scalability. We also evaluate the robustness of these bounds to changes in the number of sources. The results offer an understanding of attainable estimation accuracy of source reliability in social sensing applications that rely on un-vetted sources whose reliability is not known in advance.**

***Index Terms*—Quantification; CRLB; Scalability; Robustness; Social Sensing**

I. Introduction

Social sensing has emerged as an important paradigm of sensing applications, where humans are explicitly or implicitly involved in the process of sensing and data collection. Scalable and robust estimation of source reliability is a key challenge in social sensing due to the fact that humans are generally less reliable than well tested infrastructure sensors, and the correctness of their measurements is usually unknown *a priori*. In previous work, we developed a maximum likelihood estimator of source reliability [16] and quantified the confidence of the estimation approximately based on an *asymptotic* Cramer-Rao lower bound (CRLB) [15]. However, the asymptotic CRLB is not accurate, because it deviates significantly from the actual estimation variance in scenarios where the number of sources in the system is small. In this paper, we derive the real CRLB and show that it tracks the estimation variance tightly when the asymptotic CRLB fails to be accurate. We study the scalability limitations of real and asymptotic CRLBs and examine the robustness of the estimation performance and corresponding bounds to changes in the number of sources in the system.

It is shown that the estimation confidence can be quantified accurately. The derived real CRLB is able to characterize the estimation performance correctly when the number of sources in the system is small. Additionally, the estimation performance and the accuracy of CRLBs are shown to be robust to changes in the number of sources.

The results of this paper are important because they allow social sensing applications to assess the quality of data obtained from human participants to a desired confidence level, in the absence of independent means to verify the data and in the absence of prior knowledge of reliability of sources. This is attained via a well-founded analytic problem formulation and a solution that leverages well-known results in estimation theory.

The rest of this paper is organized as follows: We review related work in Section II. In Section III, we briefly go over the maximum likelihood estimation (MLE) approach and the problem of quantifying source reliability in social sensing applications. We then derive the real CRLB and outline the asymptotic CRLB and the confidence interval on source reliability in Section IV. The evaluation results are presented in Section V. We discuss the limitations of our model and possible extensions for future work in Section VI. Finally, we conclude the paper in Section VII.

II. Related Work

Social sensing has emerged as a new paradigm of sensing applications due to the great increase in the number of mobile sensors owned by common individuals and the proliferation of Internet connectivity. A relevant body of work, called *fact-finders*, in the machine learning and data mining communities performs trust analysis to assess the credibility of

sources and assertions claimed in information networks. Fact-finding in social sensing is more challenging because of the unknown reliability of data sources and the highly dynamic nature of social sensing topologies [1]. The basic fact-finders include Hubs and Authorities [10], Average.Log [12], and TruthFinder [17]. Other extended fact-finders further analyze properties or dependencies within assertions and sources [2], [5]–[8]. Trust analysis has been performed for both homogeneous and heterogeneous network topology [13], [18].

The Bayesian Interpretation scheme [14] represents a recent effort to convert the ranking outputs from fact-finders into the Bayesian probability semantics. However, the accuracy of truth estimation of this scheme is very sensitive to the initial conditions of iterations due to its linear approximation assumption. To overcome such limitations, Wang et al proposed a maximum likelihood estimator based on Expectation Maximization (EM) [16]. The maximum likelihood estimator provides an optimal hypothesis on source reliability and reported measurements, which is most consistent with the observed data in social sensing. The EM scheme was shown to beat Bayesian Interpretation and other state-of-art fact-finders in the estimation performance. To quantify the estimation accuracy in EM, a confidence bound based on asymptotic CRLB was proposed [15]. However, this asymptotic bound fails to be tight when the number of sources is small. In this paper, we derived, for the first time, the real CRLB that is able to accurately characterize the estimation performance for the sensing network of a small number of sources.

In statistics and estimation theory, the Cramer-Rao lower bound (CRLB) is defined as the inverse of Fisher information matrix and represents a lower bound on the estimation variance of a deterministic parameter [4]. The maximum likelihood estimation (MLE) possesses many nice asymptotic properties, one of which is called asymptotic normality. The asymptotic normality basically states that the maximum likelihood estimation is asymptotically distributed with Gaussian behavior as the data sample size increases, and the covariance of the MLE reaches the Cramer-Rao lower bound. In this paper, the confidence interval is derived by computing the CRLB of the estimation parameters and leveraging the asymptotic normality of the maximum likelihood estimation.

III. Problem Statement

For our social sensing problem, we adopt the model used for maximum likelihood source reliability estimation in social sensing [15]. Consider a social sensing application model where a group of M sources, $S_1, ..., S_M$, make individual observations about a set of N measured variables $C_1, ..., C_N$ in their environment. For example, a group of local residents might join a geo-tagging campaign to report litter locations in the park. Hence, each measured variable denotes the existence or lack thereof of litter at a given location. We consider only binary variables 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 measured variable (e.g., litter found) is encountered. Each source generally observes only a small subset of all variables (e.g., states of places they have been to).

Let us also define some notations we used: S_i denotes the i^{th} source, C_j denotes the j^{th} measured variable and S_iC_j denotes S_i reporting C_j to be true. The social sensing topology describing *who report what* can be represented by an *observation matrix* SC, where $S_iC_j = 1$ when source S_i reports that C_j is true, and $S_iC_j = 0$ otherwise. Moreover, let $P(C_j^t)$ and $P(C_j^f)$ denote the odds that the actual variable C_j is indeed true and false, respectively. Let the probability that source S_i reports an observation be s_i. Further, let the probability that source S_i is right be t_i and the probability that she is wrong be $1 - t_i$. 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 us also define a_i as the (unknown) probability that source S_i reports a variable to be true when it is indeed true, and b_i as the (unknown) probability that source S_i reports a variable to be true when it is in reality false. Formally, a_i and b_i are defined as follows:

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

The Bayes' theorem provides us with the 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 measured variable is true. Note that, this value can be known from past statistics. It does not indicate, however, whether any particular claim about a specific measured variable is true or not.

To handle the unknown correctness of measured variables in the model, a hidden variable Z is incorporated for each variable to indicate whether it is true or not (i.e., z_j is 1 when the measured variable C_j is true and 0 otherwise). A maximum-likelihood estimator [16] can now take the observation matrix SC as the input and iterate between the E-step and M-step of EM scheme 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 $\theta = (a_1, a_2, ...a_M; b_1, b_2, ...b_M)$. Our goal in this paper is to i) derive the real CRLB that accurately characterizes the estimation performance of the MLE on source reliability when the number of sources is small; ii) study the scalability limitations of both real and asymptotic CRLB; iii) evaluate the robustness of estimation performance and the derived CRLBs to changes in the number of sources.

IV. Confidence Interval Derivation from CRLB

In this section, we show that the confidence interval on source reliability is derived by computing the Cramer-Rao lower bound (CRLB) for the estimation parameters (i.e., θ) and leveraging the asymptotic normality of maximum likelihood estimation. We start with the real CRLB derivation and identify

its scalability limitation. We then outline the asymptotic CRLB that works for the sensing topology with a large number of sources. [1] Finally, we compute the confidence interval on source reliability based on the derived CRLB.

A. Real Cramer Rao Lower Bound

We first derive the real CRLB that characterizes the estimation performance 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 [9]. The estimator that reaches this lower bound is said to be *efficient*. For notational convenience, we denote the observation matrix 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 [16]:

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

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

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

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 $i = 1, 2...M$ and 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 measured variable C_j to be true given the observation matrix. After further simplification as shown in the appendix A, A_{kl} can be expressed as the summation of only the expectation terms where $j = q$:

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

where

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

We rewrite $A_{k,l} = N \bar{A}_{k,l}$ where $\bar{A}_{kl}$ is:

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

[1] The asymptotic bound was previously published in a workshop paper [15]. The workshop has no printed proceedings. This paper extends the workshop results

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 for $X_{ij} \quad i = 1, 2, ...M$ and a given j. This is much smaller than the 2^{MN} permutations for $\mathcal{X}$.

This gives us the real CRLB. Note that more measured variables 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 real 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 real CRLB when the number of sources becomes large. In this subsection, we outline the asymptotic CRLB for efficient computation in the sensing topology with a large number of sources. The asymptotic CRLB is derived based on the assumption that the correctness of the hidden variable (i.e., z_j) can be correctly estimated from EM. This is a reasonable assumption when the number of sources is sufficient [15]. Under this assumption, the log-likelihood function of the maximum likelihood estimation we get from EM can be expressed as follows:

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

We first compute the Fisher Information Matrix at the MLE from the log-likelihood function given by Equation (17). According to prior work [16], the maximum likelihood estimator $\hat{\theta}_{MLE}$ is given by:

$$\hat{a}_i^{MLE} = \frac{\sum_{j \in SJ_i} Z_j^c}{\sum_{j=1}^{N} Z_j^c} \quad \hat{b}_i^{MLE} = \frac{K_i - \sum_{j \in SJ_i} Z_j^c}{N - \sum_{j=1}^{N} Z_j^c} \tag{18}$$

where SJ_i is the set of measured variables reported by source S_i and Z_j^c is the converged probability of the j^{th} measured variable 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., measured variables).

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 measured variables as:

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

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 [9]. 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 norm 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 topology 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 topology 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).[2] Therefore, $(\hat{t}_i^{MLE} - t_i^0)$ also follows a norm distribution with 0 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.

In this section, we derived a confidence interval that allows social sensing applications to assess the accuracy of their estimation of reliability of sources. Hence, applications can not only produce a best hypothesis regarding correctness of sources, but also compute their confidence in this hypothesis. In the following section, we evaluate the accuracy of the computed confidence bounds.

V. Evaluation

In this section, we present the evaluation of the performance of the computed confidence interval of source reliability and the derived CRLBs in social sensing. We built a simulator in Matlab 7.10.0 that generates a random number of sources and measured variables. 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). For each source S_i, L_i observations are generated. Each observation has a probability P_i of being true (i.e., reporting a variable as true correctly) and a probability $1 - P_i$ of being false (reporting a variable as true when it is not). One can think of these variables as observed "problems". Sources do not report "lack of problems". Hence, they never report a variable to be false. We let P_i be uniformly distributed between 0.5 and 1 in our experiments[3].

A. Evaluation of Confidence Interval

In this subsection, we evaluate the performance of the confidence interval on source reliability derived in the previous section. We carried out experiments over three different observation matrix scales: small, medium and large. The simulation parameters are listed in Table I. The total number of measured variables is the sum of both true and false ones. The average observations reported by each source is set to 100. For each observation matrix scale, we run the EM algorithm and compute the confidence interval on source reliability based on Equation (23). We repeat the experiments 100 times for each observation matrix scale. Three representative confidence levels (i.e., 68%, 90%, 95%) are used in our evaluation.

Observation Matrix Scale	Number of Sources	Number of True Measured Variables	Number of False Measured Variables
Small	100	500	500
Medium	200	1000	1000
Large	300	2000	2000

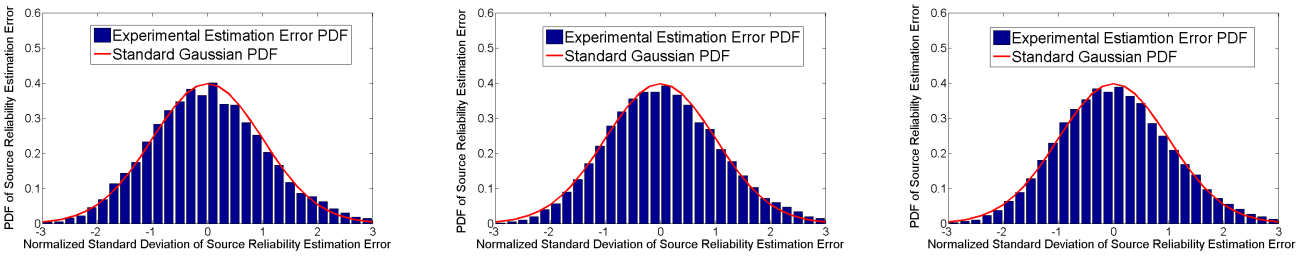
TABLE I
Parameters of Three Typical Observation Matrix Scale

Figure 1 shows the normalized probability density function (PDF) of source reliability estimation error over three observation matrix scales. We computed the experimental PDF by leveraging the actual estimation error (i.e., compare to the ground truth) and the confidence interval derived in Section IV. We compared the experimental PDF with the standard Gaussian distribution to verify the asymptotic normality property of estimation results. We observe the experimental PDF match well with the theoretical Gaussian distribution over three observation matrix scales.

Figure 2 shows the comparison between the actual estimation confidence and three different confidence levels we set for the small observation matrix scenario. The actual estimation confidence is computed as the percentage of sources whose estimation error stay within the corresponding confidence bound for every experiment. This percentage represents the probability that a randomly chosen source keeps its reliability estimation error within the confidence bound. We observe that the actual estimation confidence of using 3 different confidence bounds stays close to the corresponding confidence levels we used for the experiment. Moreover, at higher confidence levels, a lower fluctuation of the actual estimation confidence is observed. Similar results are observed for the medium and large observation matrices as well, which are shown in Figure 3 and Figure 4. Additionally, we also note that the fluctuation of the actual estimation confidence decreases as the observation matrix scale increases. This is because the estimation variance

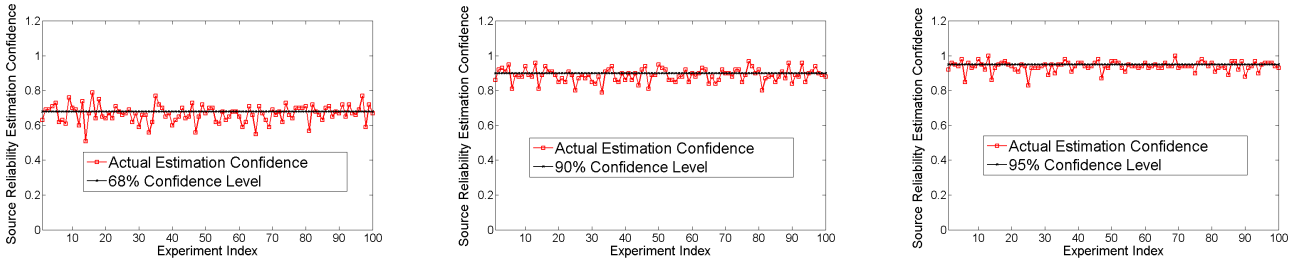
[2]The value of s_i can be estimated as $\frac{L_i}{N}$, where L_i is the number of observations reported by source S_i

[3]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



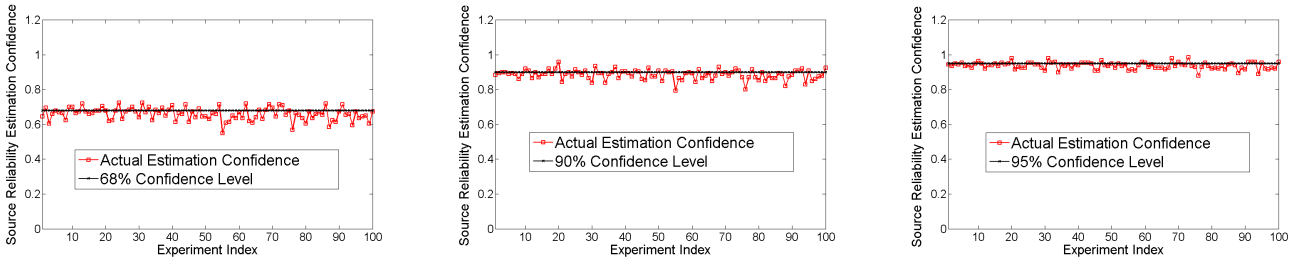
(a) Small Observation Matrix (b) Medium Observation Matrix (c) Large Observation Matrix

Fig. 1. Normalized Source Reliability Estimation Error PDF



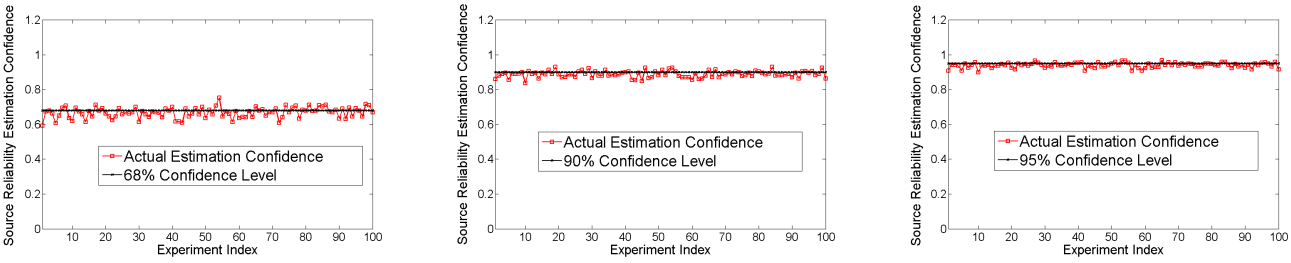
(a) 68% Confidence Level (b) 90% Confidence Level (c) 95% Confidence Level

Fig. 2. Source Reliability Estimation Confidence for Small Observation Matrix



(a) 68% Confidence Level (b) 90% Confidence Level (c) 95% Confidence Level

Fig. 3. Source Reliability Estimation Confidence for Medium Observation Matrix



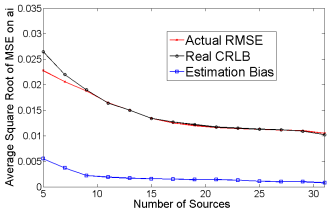
(a) 68% Confidence Level (b) 90% Confidence Level (c) 95% Confidence Level

Fig. 4. Source Reliability Estimation Confidence for Large Observation Matrix

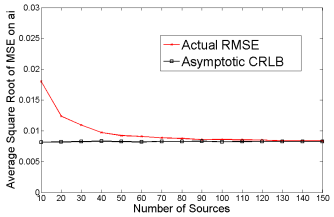
characterized by CRLB is inversely proportional to the number of measured variables in the system, which will be further evaluated in the next subsection.

B. Evaluation of CRLB

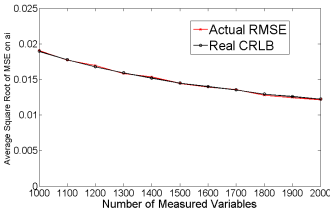
In this subsection, we evaluate the performance of derived CRLBs (both real and asymptotic) by comparing them to the actual estimation variance of the estimation parameter (i.e., a_i, b_i). The actual estimation variance is characterized by the average RMSE (square root of the mean squared error) of all sources. The first experiment evaluates the effect of the number of sources (i.e., M) in the system on the CRLB performance. We start with the real CRLB evaluation. We fix the true and false measured variables to be 1000 respectively, the average observations per source is set to 100. We vary the number of sources from 5 to 31. Reported results are averaged over 100 experiments and are shown in Figure 5. Observe that the real CRLB tracks the actual estimation 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 Real CRLB when there are too few sources. This is because the MLE is biased on those points



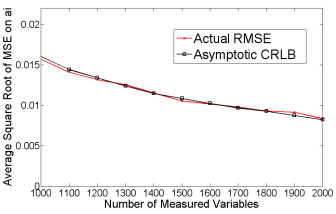
(a) Real CRLB of a_i



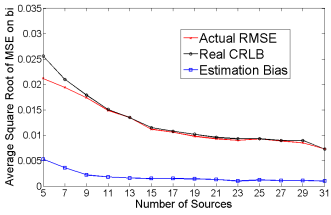
(a) Asymptotic CRLB of a_i



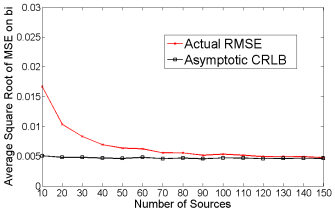
(a) Real CRLB of a_i



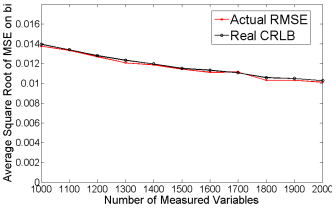
(a) Asymptotic CRLB of a_i



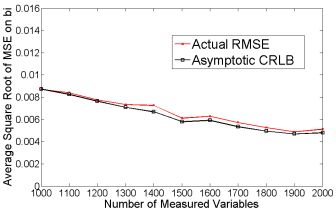
(b) Real CRLB of b_i



(b) Asymptotic CRLB of b_i



(b) Real CRLB of b_i



(b) Asymptotic CRLB of b_i

Fig. 5. Real CRLB of a_i and b_i versus Vaying M

Fig. 6. Asymptotic CRLB of a_i and b_i versus Vaying M

Fig. 7. Real CRLB of a_i and b_i versus Vaying N

Fig. 8. Asymptotic CRLB of a_i and b_i versus Vaying N

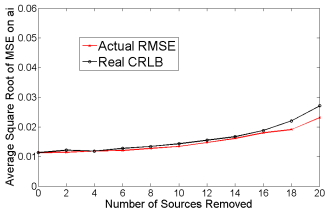
due to the small dataset. As illustrated in Section IV-A, the computation of real CRLB does not scale with the number of sources in the system. Hence, we also evaluate the the performance of asymptotic CRLB when the number of sources becomes large. We keep the experimental configuration the same as above, but change the number of sources from 10 to 150. Results are shown in Figure 6. We observe that the asymptotic CRLB deviates 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 actual RMSE converges to the asymptotic CRLB quickly and the difference between the two becomes insignificant.

The second experiment compares the derived CRLBs (both real and asymptotic) to the actual RMSE of estimation parameters when the number of measured variables (i.e., N) changes. As shown in Section IV, both the real and asymptotic CRLB decrease as $\frac{1}{N}$. As before, we first evaluate the performance of real CRLB. We fix the number of sources as 20, the average number of observations per source is set to 100. We also keep the number of true and false measured variables the same. The number of measured variables varies from 1000 to 2000. Reported results are averaged over 100 experiments and are shown in Figure 7. We observe that the real CRLB is able to track the actual RMSE on estimation parameter correctly and they both decrease approximately as $\frac{1}{N}$ when the number of measured variable increases. Similarly, we carry out the experiment to evaluate the performance of asymptotic CRLB. We keep the experimental configuration the same as above, but set the number of sources to be 150. Results are shown in Figure 8. We observe that the asymptotic CRLB also follows closely on the actual RMSE of the estimation parameter and they reduce approximately as $\frac{1}{N}$ when the number of measured variable increases.

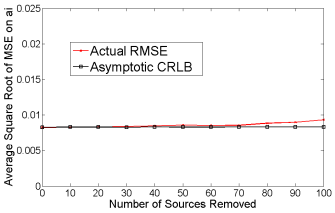
C. Sensitivity to Changing the Number of Sources

In this subsection, we evaluate the robustness (or sensitivity) of the estimation performance and the derived CRLBs when the number of sources changes under different source reliability distributions. The key characteristic that determines the resilience of a network is the network topology. The social sensing topology is characterized by the link connections between sources and two sets of measured variables (i.e., true and false). The link connection skew is mainly determined by the source reliability distribution. We consider two representative network topologies: scale-free and exponential topologies in our evaluation. For scale-free topology, sources have diverse reliability and nodes with high reliability form the "hubs" of the network. For exponential topology, sources have similar reliability and nodes with higher reliability are exponentially less probable. Our experiments were done by source removal (i.e., sources are randomly selected and removed from the system). This represents the scenario where random sources decide to quit the sensing application or their sensing devices fail. However, it is equivalent to reversing the steps and investigating the addition of sources.

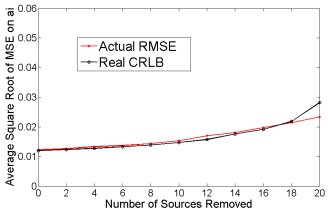
In the first experiment, we evaluate the estimation performance and the derived CRLBs of the scale-free network topology. To generate the scale-free network topology, we let the source reliability follow a uniform distribution on its definition range. We first evaluate the performance of the real CRLB compared to the actual RMSE on the estimation parameter. We fix both the number of true and false measured variables to 1000. The average number of observations per source is set to 100. We start with 25 sources and gradually remove sources from the system. Figure 9 shows the real CRLB and actual RMSE of the estimation parameter. Observe that the estimation performance (i.e., actual RMSE) degrades gracefully and the real CRLB tracks the actual RMSE reasonably well as the number of removed sources increases. Also note that the real CRLB deviates slightly from the RMSE when majority of sources are removed from the system. We then repeat similar experiments for the asymptotic CRLB as well. We start with 150 sources and gradually remove the sources from the system.



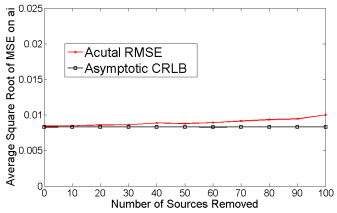
(a) Real CRLB of a_i



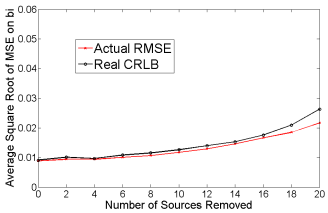
(a) Asymptotic CRLB of a_i



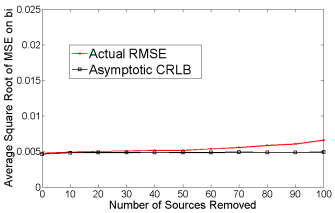
(a) Real CRLB of a_i



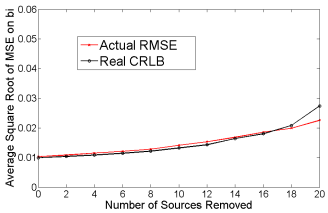
(a) Asymptotic CRLB of a_i



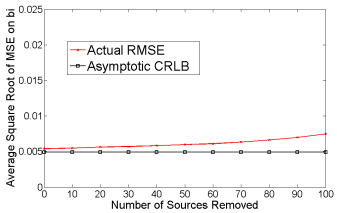
(b) Real CRLB of b_i



(b) Asymptotic CRLB of b_i



(b) Real CRLB of b_i



(b) Asymptotic CRLB of b_i

Fig. 9. Real CRLB of a_i and b_i versus Source Removal of Scale-free Topology

Fig. 10. Asymptotic CRLB of a_i and b_i versus Source Removal of Scale-free Topology

Fig. 11. Real CRLB of a_i and b_i versus Source Removal of Exponential Topology

Fig. 12. Asymptotic CRLB of a_i and b_i versus Source Removal of Exponential Topology

Results are shown in Figure 10. The results for asymptotic CRLB are similar to real CRLB.

In the second experiment, we evaluate the estimation performance and the derived CRLBs of the exponential network topology. To generate the exponential network topology, we let the source reliability follow a norm distribution (with the mean value as the mean of its definition range and a reasonably small variance). As we did before, we first evaluate the performance of the real CRLB compared to the actual RMSE on the estimation parameter. The standard deviation of the norm distribution of source reliability is set to 0.02, other settings are kept the same as the first experiment. Figure 11 shows the real CRLB and actual RMSE of the estimation parameter. Observe that actual RMSE increases gradually as the number of removed sources grows and the real CRLB tracks the actual RMSE well. We then repeat similar experiments for the asymptotic CRLB as well. The experimental settings are kept the same as the first experiment. Results are shown in Figure 12. Similar results as we have for the real CRLB are observed for the asymptotic CRLB.

For both the scale-free and exponential topology of social sensing, the above results show that the estimation performance is relatively robust (or insensitive) to changes in the number of sources in the network. Both real and asymptotic CRLBs are able to track the estimation performance as long as a limited number of sources stay in the system.

VI. Limitations and Future Work

This paper studies the scalability and robustness limitations of the confidence bounds to characterize the estimation performance on source reliability 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 (i.e., copy observations, forward rumor, and etc.) or even collude to misrepresent the truth. 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 [11]. The above techniques can be used together with our quantification scheme to handle source dependency. Moreover, authors are also working on extending the current model to handle non-independent sources. For example, one could cluster dependent sources into approximately independent ones according to some source similarity metric and run our scheme on top of the clustered sources. Additionally, sources are sometimes experts in specific domains. It would be interesting to assess the estimation performance on source reliability by taking source expertise into consideration. One possibility is to weight observations differently depending on the source's expertise in the confidence calculation.

No dependencies were assumed among different measured variables. There may be cases, however, observations on one measured variable could imply observations on another (e.g., "flooding" at city B may imply "raining" at city A). The background knowledge of the observation dependency can thus be integrated with our scheme to pre-process the observation matrix (e.g., add or remove links) based on the reported observations and their relationship. 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, the source reliability and confidence estimation will 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 high reliability and confidence in estimation by simply making many trivially true observations. There are techniques that analyze the hardness of observations, which is possible to be integrated with our scheme [8]. In this paper, sources are assumed to report

positive states of measured variables (e.g., litter found) only and ignore the negative states. This is a reasonable assumption for some typical social sensing applications (e.g., geotagging). However, sources can also make contradicting observations in other types of applications (e.g., on-line review system). Our model can be extended to handle contradicting observations by expanding the estimation parameter vector that covers only positive states to both positive and negative states and rebuilding the likelihood function. The general outline of the proof still holds true in this scenario.

Having the fundamental estimation error analysis and confidence quantification theory in place, we can relax the above assumptions and accommodate the mentioned extensions in future work. The authors are currently working on the above extensions.

VII. Conclusion

This paper presents new confidence bounds on source reliability in social sensing applications that allow the applications to accurately assess the quality of data contributed by human participants to a desired confidence level. The confidence bounds are computed based on the Cramer-Rao lower bound (CRLB) of the maximum likelihood estimation of source reliability. The real and asymptotic CRLBs are derived and their scalability limitations are examined. The estimation performance and accuracy of the derived CRLBs are shown to be robust to changes in the number of sources for different sensing network topologies.

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:

$$
\begin{aligned}
&E\Big[\frac{(2X_{kj}-1)Z_j(2X_{lq}-1)Z_q}{a_k^{X_{kj}}(1-a_k)^{(1-X_{kj})}a_l^{X_{lq}}(1-a_l)^{(1-X_{lq})}}\Big] = \\
&\sum_{x\in\mathcal{X}}(2X_{kj}-1)(2X_{lq}-1)\times \\
&\Big(\prod_{\substack{i=1\\ i\neq k}}^{M} a_i^{X_{ij}}(1-a_i)^{(1-X_{ij})}\times d\prod_{\substack{i=1\\ i\neq l}}^{M} a_i^{X_{iq}}(1-a_i)^{(1-X_{iq})}\times d\Big) \\
&\times\Bigg(\prod_{\substack{j'=1\\ j'\neq j \text{ or } q}}^{N}\Big\{\prod_{i=1}^{M} a_i^{X_{ij'}}(1-a_i)^{(1-X_{ij'})}\times d \\
&+\prod_{i=1}^{M} b_i^{X_{ij'}}(1-b_i)^{(1-X_{ij'})}\times(1-d)\Big\}\Bigg) \\
&=\sum_{x\in\mathcal{X}j\times\mathcal{X}_q}(2X_{kj}-1)(2X_{lq}-1)\times \\
&\Big(\prod_{\substack{i=1\\ i\neq k}}^{M} a_i^{X_{ij}}(1-a_i)^{(1-X_{ij})}\times d\prod_{\substack{i=1\\ i\neq l}}^{M} a_i^{X_{iq}}(1-a_i)^{(1-X_{iq})}\times d\Big) \\
&=\sum_{X_{kj}=0}^{1}\sum_{X_{lq}=0}^{1}(2X_{kj}-1)(2X_{lq}-1)=0 \qquad j\neq q \quad (24)
\end{aligned}
$$

Acknowledgements

Research reported in this paper was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. This work was partially supported by the LCCC and eLLIIT centers at Lund University, Sweden. 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.

References

[1] C. Aggarwal and T. Abdelzaher. Integrating sensors and social networks. *Social Network Data Analytics, Springer*, expected in 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] X. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava. Global detection of complex copying relationships between sources. *PVLDB*, 3(1):1358–1369, 2010.

[6] X. Dong, L. Berti-Equille, and D. Srivastava. Truth discovery and copying detection in a dynamic world. *VLDB*, 2(1):562–573, 2009.

[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. V. Hogg and A. T. Craig. *Introduction to mathematical statistics*. Prentice Hall, 1995.

[10] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[11] 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 *Proceedings of the 27th International Conference on Distributed Computing Systems*, ICDCS '07, pages 56–, Washington, DC, USA, 2007. IEEE Computer Society.

[12] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *International Conference on Computational Linguistics (COLING)*, 2010.

[13] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *15th SIGKDD international conference on Knowledge discovery and data mining (KDD'09)*, pages 797–806, 2009.

[14] 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.

[15] D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal. On quantifying the accuracy of maximum likelihood estimation of participant reliability in social sensing. In *DMSN11: 8th International Workshop on Data Management for Sensor Networks*, August 2011.

[16] 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.

[17] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE Trans. on Knowl. and Data Eng.*, 20:796–808, June 2008.

[18] X. Yin and W. Tan. Semi-supervised truth discovery. In *WWW*, New York, NY, USA, 2011. ACM.