On Quantifying the Accuracy of Maximum Likelihood Estimation of Participant Reliability in Social Sensing

Dong Wang, Tarek Abdelzaher
Department of Computer Science
University of Illinois
Urbana, IL 61801
dwang24@illinois.edu, zaher@illinois.edu

Lance Kaplan
Networked Sensing & Fusion Branch
US Army Research Laboratory
Adelphi, MD 20783
lance.m.kaplan@us.army.mil

Charu C. Aggarwal
IBM Research
Yorktown Heights, NY 10598
charu@us.ibm.com

ABSTRACT

This paper presents a confidence interval quantification of maximum likelihood estimation of participant reliability in social sensing applications. The work is motivated by the emergence of social sensing as a data collection paradigm, where humans perform the data collection tasks. A key challenge in social sensing applications lies in the uncertain nature of human measurements. Unlike well-calibrated and well-tested infrastructure sensors, humans are less reliable, and the likelihood that participants' measurements are correct is often unknown a priori. Hence, it is hard to estimate the accuracy of conclusions made based on social sensing data. In previous work, we developed a maximum likelihood estimator of reliability of both participants and facts concluded from the data. This paper presents an analytically-founded bound that quantifies the accuracy of such maximum likelihood estimation in social sensing. A confidence interval is derived by leveraging the asymptotic normality of maximum likelihood estimation and computing the approximation of Cramer-Rao bound (CRB) for the estimation parameters. The proposed quantification approach is empirically validated and shown to accurately bound the actual estimation error given sufficient number of participants under different sensing topologies.

1. INTRODUCTION

Social sensing, where individuals act as sensors, is a key emerging category of sensing applications. Yet, quantifying the reliability of data collected from human sources remains one of the main challenges in utilizing social sensing in mission-critical systems. This reliability problem has long been known in military scenarios and is becoming increasingly important in commercial and civil settings as well. The main research question is one of quantifying a level of confidence in information reported by a group of human observers.

This paper quantifies confidence in social sensing observations by computing a confidence interval of a maximum likelihood estimator of participant reliability from social sensing data. The confidence interval is computed based on the *approximation* of Cramer-Rao bound (CRB) [4], which quantifies the variance of a minimum-variance unbiased estimator. This bound approximates the CRB by leveraging observations from participants and knowledge of truthfulness of facts estimated from an maximum likelihood estimator of social sensing [14].

We consider a sensing application in which data are collected from a large population, where the reliability (i.e., the probability of correctness) of individual participants, and hence observations made by them, is not known *a priori*. We aim to derive the confidence interval in the maximum likelihood estimation of participant and observation reliability given no prior knowledge other than the information describing who reported which observations. The maximum likelihood estimation problem itself is not the topic of this paper. An approach based on expectation maximization [5] is described in prior work [14]. This paper quantifies the *confidence* approximately in the answer reported by such an estimator.

We concern ourselves with binary measurements only; for example, reporting whether or not a given person or object was seen at a given location. Our derivation leverages the asymptotic normality of maximum likelihood estimation and computes the approximation of Cramer-Rao bound (CRB) of the estimation parameters used in an expectation maximization scheme. It is shown that the probability of the estimated participant reliability falling into the derived confidence interval is always greater than the given confidence level, as long as enough participants report sufficient observations, and as long as some participants make the same observation. In other words, it is shown that our confidence window correctly bounds estimation error.

The rest of this paper is organized as follows: In Section 2, we review a maximum likelihood estimation approach for social sensing applications and formulate the problem of deriving the confidence interval of participant reliability. The derivation of the proposed approximation for CRB quantification on maximum likelihood estimation is discussed in Section 3. Evaluation results are presented in Section 4. We

8th International Workshop on Data Management for Sensor Networks (DMSN 2011)

review related work in Section 5. Finally, we conclude the paper in Section 6.

2. PROBLEM FORMULATION

First, we review the maximum likelihood estimation approach for truth discovery in social sensing applications. We consider a social sensing application model where a group of M participants, $S_1, ..., S_M$, make individual observations about a set of N measured variables $C_1, ..., C_N$ in their environment. Each measured variable denotes the existence or lack thereof of certain phenomenon of application's interests. In this effort, we consider only binary variables and assume, without loss of generality, that their "normal" state is negative. Hence, participants report only when a positive value is encountered. Each participant generally observes only a subset of all variables. Our goal is to determine the confidence interval of the participant reliability maximum likelihood estimation for a given confidence level based only on the information of which observations are reported by which participant.

Let S_iC_j denote an observation reported by participant S_i claiming that C_j is true. Let $P(C_j^t)$ and $P(C_j^f)$ denote the probability that the actual variable C_j is indeed true and false, respectively. Different participants may make different numbers of observations. Let the probability that participant S_i makes an observation be s_i. Further, let the probability that participant S_i is right be t_i and the probability that it is wrong be $1 - t_i$. Note that, this probability depends on the participant'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 participant S_i reports a variable to be true when it is indeed true, and b_i as the (unknown) probability that participant S_i reports a variable to be true when it is in reality false. Formally, a_i and b_i are defined as follows:

$$\begin{aligned} a_i &= P(S_iC_j | C_j^t) \\ b_i &= P(S_iC_j | C_j^f) \end{aligned} \tag{2}$$

From the definition of t_i, a_i and b_i, we can determine their relationship using the Bayesian theorem:

$$\begin{aligned} a_i = P(S_iC_j|C_j^t) = \frac{P(S_iC_j, C_j^t)}{P(C_j^t)} = \frac{P(C_j^t|S_iC_j)P(S_iC_j)}{P(C_j^t)} \\ b_i = P(S_iC_j|C_j^f) = \frac{P(S_iC_j, C_j^f)}{P(C_j^f)} = \frac{P(C_j^f|S_iC_j)P(S_iC_j)}{P(C_j^f)} \end{aligned} \tag{3}$$

The key input to the maximum likelihood estimator algorithm is a matrix SC, where $S_iC_j = 1$ when participant S_i reports that C_j is true, and $S_iC_j = 0$ otherwise. Let us call it the *observation matrix*. For initialization, we also define the background bias d to be 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 measured variable is true or not. To initialize the algorithm, we set $P(C_j^t) = d$ and set $P(S_iC_j) = s_i$. Plugging these, together with t_i into the definition of a_i and b_i, we get the initial values:

$$\begin{aligned} a_i &= \frac{t_i \times s_i}{d} \\ b_i &= \frac{(1 - t_i) \times s_i}{1 - d} \end{aligned} \tag{4}$$

The best (in the sense of maximum likelihood) estimate $\hat{t}_i^*$ of the reliability of each participant S_i can be obtained by using the Expectation Maximization (EM) algorithm [5]. In the EM algorithm, a latent variable Z is introduced for each measured 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). The observation matrix SC is treated as the observed data X, and $\theta = (a_1, a_2, ...a_M; b_1, b_2, ...b_M)$ is the parameter vector of the model we want to estimate. The EM algorithm iteratively performs an expectation step (E-step) and a maximization step (M-step) to compute the best estimate of the parameter θ that maximizes the expected logarithm likelihood function. The likelihood function used by EM scheme is given by:

$$\begin{aligned} &L(\theta; X, Z) = p(X, Z|\theta) \\ &= \prod_{j=1}^{N} \left\{ \prod_{i=1}^{M} a_i^{S_iC_j} (1 - a_i)^{(1 - S_iC_j)} \times d \times z_j \right. \\ &\left. + \prod_{i=1}^{M} b_i^{S_iC_j} (1 - b_i)^{(1 - S_iC_j)} \times (1 - d) \times (1 - z_j) \right\} \end{aligned} \tag{5}$$

An output of the EM algorithm is the maximum likelihood estimation of each participant's reliability, which is most consistent with the observation matrix SC. However, an important problem that remains unanswered from the maximum likelihood estimation of the EM scheme is: what is the confidence interval of the resulting participant reliability estimation? Only by answering this question, can we completely characterize estimation performance, and hence participant reliability in social sensing applications. The goal of this paper is to demonstrate, in an analytically founded manner, how to compute the confidence interval of each participant's reliability. Formally, this is given by:

$$(\hat{t}_i^{MLE} - c_p^{lower}, \hat{t}_i^{MLE} + c_p^{upper}) \qquad c\% \tag{6}$$

where $c\%$ is the confidence level of the estimation interval, c_p^{lower} and c_p^{upper} represent the lower and upper bound on the estimation deviation from the maximum likelihood estimation $\hat{t}_i^{MLE}$ respetively. We target to find c_p^{lower} and c_p^{upper} for a given $c\%$ and an observation Matrix SC.

3. RELIABILITY DERIVATION

In this section, we derive a confidence interval based on the approximation of Cramer-Rao Bound and the aforementioned formulation of maximum likelihood estimation of participant reliability. The log-likelihood function (or log-probability density function) of the maximum likelihood estimation we get from EM can be expressed as:

$$\begin{aligned} &l_{em}(x; \theta) = \log p_{em}(x; \theta) \\ &= \sum_{j=1}^{N} \left\{ z_j \times \left[\sum_{i=1}^{M} (S_iC_j \log a_i + (1 - S_iC_j) \log(1 - a_i) + \log d) \right] \right. \\ &+ (1 - z_j) \\ &\left. \times \left[\sum_{i=1}^{M} (S_iC_j \log b_i + (1 - S_iC_j) \log(1 - b_i) + \log(1 - d)) \right] \right\} \end{aligned} \tag{7}$$

The likelihood (or probability density function) is:

$$p_{em}(x;\theta) = \exp(l_{em}(x;\theta)) \tag{8}$$

The goal here is to show that the confidence interval of the estimated parameter θ can be asymptotically characterized by the approximation of Cramer-Rao bound (CRB) given the observation matrix as well as estimated truthfulness of each measured variable from EM scheme.

In statistic mathematics and 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 partial derivative of the log-likelihood function $l(x;\theta)$ with respect to θ is called the score. A score vector $\psi(x;\theta)$ for a $k \times 1$ estimation vector $\theta = [\theta_1, \theta_2, ..., \theta_k]^T$ is defined as:

$$\psi(x;\theta) = [\frac{\partial l(x;\theta)}{\partial \theta_1}, \frac{\partial l(x;\theta)}{\partial \theta_2}, ..., \frac{\partial l(x;\theta)}{\partial \theta_k}]^T \tag{9}$$

The Fisher information is defined as the second moment of the score vector:

$$I(\theta) = E_X[\psi(X;\theta)\psi(x;\theta)^T] \tag{10}$$

where the expectation is taken over all values for X with respect to the probability function $p(x;\theta)$ for any given value of θ.

Hence, the Fisher information for the above estimation vector θ takes the form of an $k \times k$ matrix, the Fisher Information Matrix, with the representative element:

$$(I(\theta))_{i,j} = E_X[(\frac{\partial l(x;\theta)}{\partial \theta_i})(\frac{\partial l(x;\theta)}{\partial \theta_j})] \tag{11}$$

If $l(x;\theta)$ is twice differentiable with respect to θ (which happens to be the case for EM model), under certain regularity conditions, the Fisher Information Matrix may also be written as [10]:

$$(I(\theta))_{i,j} = -E_X[\frac{\partial^2 l(x;\theta)}{\partial \theta_i \partial \theta_j}] \tag{12}$$

In estimation theory and statistics, the Cramer-Rao bound (CRB) expresses a lower bound on the variance of estimators of a deterministic parameter. 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*.

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]:

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N(0, I^{-1}(\hat{\theta}_{MLE})) \tag{13}$$

where n is the sample size, 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 to estimate its true 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 bound (i.e., an efficient estimator).

Since the estimator we obtain from the EM algorithm is a maximum likelihood estimator of the parameter θ, we now show how to leverage the asymptotic normality and the approximation of Cramer-Rao bound to derive a confidence interval that quantifies the estimation accuracy of the estimated parameter θ for the model of the EM scheme.

We first compute the approximation of Fisher Information Matrix from the log-likelihood function given by Equation (7). Note that this computation utilizes the estimated truthfulness of each measured variable from EM scheme, hence offers approximated results. According to prior work [14], the maximum likelihood estimator $\hat{\theta}_{MLE}$ is given by:

$$\begin{aligned} \hat{a}_i^{MLE} &= \frac{\sum_{j \in SJ_i} Z_j^c}{\sum_{j=1}^{N} Z_j^c} \\ \hat{b}_i^{MLE} &= \frac{K_i - \sum_{j \in SJ_i} Z_j^c}{N - \sum_{j=1}^{N} Z_j^c} \end{aligned} \tag{14}$$

where SJ_i is the set of measured variables reported by participant S_i and Z_j^c is the converged value of $Z(t,j)$ (i.e., $p(z_j = 1|X_j, \theta^{(t)})$) 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). The Fisher information in a random sample of size n is n times the Fisher information in one observation [10], and hence

$$I_n(\theta) = nI(\theta) \tag{15}$$

Plugging $l_{em}(x;\theta)$ given by Equation (7) into the Fisher Information Matrix defined in Equation (12), we have:

$$(I(\hat{\theta}_{MLE}))_{i,j} \tag{16}$$
$$= \begin{cases} 0 & i \neq j \\ -E_X\left[\frac{1}{N}\frac{\partial^2 l_{em}(x;a_i)}{\partial a_i^2}|_{a_i=\hat{a}_i^{MLE}}\right] & i = j \in [1, M] \\ -E_X\left[\frac{1}{N}\frac{\partial^2 l_{em}(x;b_i)}{\partial b_i^2}|_{b_i=\hat{b}_i^{MLE}}\right] & i = j \in (M, 2M] \end{cases}$$

Observe that the Fisher Information Matrix of the maximum-likelihood estimator from the EM scheme is a *diagonal matrix*, hence the inverse of this matrix is:

$$(I^{-1}(\hat{\theta}_{MLE}))_{i,j} \tag{17}$$
$$= \begin{cases} 0 & i \neq j \\ -E_X\left[\frac{N}{\frac{\partial^2 l_{em}(x;a_i)}{\partial a_i^2}}|_{a_i=\hat{a}_i^{MLE}}\right] & i = j \in [1, M] \\ -E_X\left[\frac{N}{\frac{\partial^2 l_{em}(x;b_i)}{\partial b_i^2}}|_{b_i=\hat{b}_i^{MLE}}\right] & i = j \in (M, 2M] \end{cases}$$

From the asymptotic normality of the maximum likelihood estimator specified by Equation (13), we know that $(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N(0, \frac{1}{N} I^{-1}(\hat{\theta}_{MLE}))$. Therefore, substituting $(I^{-1}(\hat{\theta}_{MLE}))$ by Equation (17) into Equation (13), we obtain the covariance matrix $Cov(\hat{\theta}_{MLE})$ of the asymptotic normal distribution for the maximum likelihood estimation of EM scheme, which is given by:

$$(Cov(\hat{\theta}_{MLE}))_{i,j} \tag{18}$$
$$= \begin{cases} 0 & i \neq j \\ -E_X\left[\frac{1}{\frac{\partial^2 l_{em}(x;a_i)}{\partial a_i^2}}|_{a_i=\hat{a}_i^{MLE}}\right] & i = j \in [1, M] \\ -E_X\left[\frac{1}{\frac{\partial^2 l_{em}(x;b_i)}{\partial b_i^2}}|_{b_i=\hat{b}_i^{MLE}}\right] & i = j \in (M, 2M] \end{cases}$$

Using the converged log-likelihood function of Equation (7) and substituting Equation (14) into Equation (18), the above covariance matrix can be further written as:

$$(Cov(\hat{\theta}_{MLE}))_{i,j} \tag{19}$$
$$= \begin{cases} 0 & i \neq j \\ \frac{\hat{a}_i^{MLE} \times (1-\hat{a}_i^{MLE})}{N \times d} & i = j \in [1, M] \\ \frac{\hat{b}_i^{MLE} \times (1-\hat{b}_i^{MLE})}{N \times (1-d)} & i = j \in (M, 2M] \end{cases}$$

Note that, the actual CRB bound is a function of both M and N. However, the approximation CRB bound derived is independent of M. Let us denote the variance of estimation error on parameter a_i as $Var(\hat{a}_i^{MLE})$. Recall the relation between participant reliability and estimation parameter a_i is $a_i = \frac{t_i \times s_i}{d}$. For a given topology, s_i and d are known constants, $(\hat{t}_i^{MLE} - t_i^0)$ also follows a norm distribution with 0 mean and variance given by:

$$Var(\hat{t}_i^{MLE}) = \left(\frac{d}{s_i}\right)^2 Var(\hat{a}_i^{MLE}) \tag{20}$$

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

$$(\hat{t}_i^{MLE} - c_p\sqrt{Var(\hat{t}_i^{MLE})}, \hat{t}_i^{MLE} + c_p\sqrt{Var(\hat{t}_i^{MLE})}) \tag{21}$$

where c_p is the standard score (z-score) of the confidence level p. For example, for the 95% confidence level, $c_p = 1.96$. Note that the derived confidence interval of the participant reliability maximum likelihood estimator can be computed by simply using the converged maximum lieklihood estimation of the EM scheme. This completes the deriviation.

4. EVALUATION

In this section, we carry out simulation experiments to evaluate the performance of the computed confidence interval of participant reliability in social sensing. We built a simulator in Matlab 7.10.0 that generates a random number of participants and measured variables. A random probability P_i is assigned to each participant S_i representing his/her reliability (i.e., the ground truth probability that they report correct observations). For each participant 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". Participants 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[1].

We evaluate the derived confidence interval on participant reliability over three different observation matrix scales: small, medium and large. The simulation parameters of the three observation matrix scales are listed in Table 1. The average observations reported by each participant is set to 100. For each observation matrix scale, we run the EM algorithm and compute the confidence interval on participant reliability based on Equation (21). We repeat the experiments 100 times for each observation matrix scale and call the experiments with the actual estimation error falling outside the confidence interval *outliers*. We choose three representative confidence levels (i.e., 68%, 90%, 95% [2]), respectively. For a given confidence level, we further define the participant who has the largest number of outliers over all experiments as the *worst-case participant*. Hence, we record the number of ourliters of every worst-case particpant for a given confidence level and compare it with the theoretical maximum number of outliers.

Observation Matrix Scale	Number of Participants	Number of True Measured Variables	Number of False Measured Variables
Small	50	200	200
Medium	100	500	500
Large	200	1000	1000

Table 1: Parameters of Three Typical Observation Matrix Scale

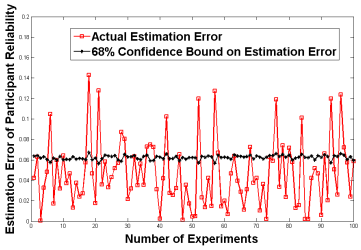
Figure 1 shows the confidence interval bounds on the participant reliability estimation error with three different confidence levels for the small observation matrix. Note that the CRB is simply a function of the ground truth parameter values. However, it is reasonable to substitute the true (but unknown) parameter values with their ML estimates [10]. This is the reason that bounds in the figure appear to fluctuate rather than being flat. Observe that the actual estimation error on participant reliability is well bounded by its corresponding confidence interval. Specially, the numbers of outliers for the worst-case participant at confidence levels 68%, 90% and 95% are 28, 7 and 4 out of 100 experiments. They are less than the theorectial maixmum number of outliers for the three confidence levels (i.e., 32, 10 and 5 out of 100 experiments, predicted using our derived bound). Similar results are observed for the medium and large observation matrices as well, which are shown in Figure 2 and Figure 3.

A summary of the comparison between confidence interval bounds in estimating participant reliability and the theoretical results is shown in Table 2. We observe that the probability of the estimated participant reliability falling into the derived confidence interval is always greater than the corresponding confidence level.

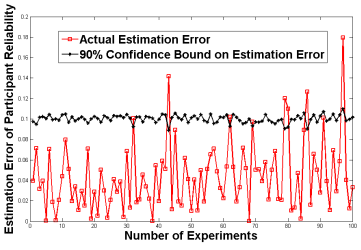
Since the derived bound is an approximation of true CRB (i.e., it depends on the correct estimation of truthfulness of measured variables from EM scheme), we study the conditions when such approximated bound fails to bound the actual error of the estimation parameters. We fix the true and false measured variables to be 1000 respectively, the average observations per participant is set to 100. We vary the number of participants from very small (i.e., 5) to large (i.e., 205). Reported results are averaged over 100 experiments. Figure 4 shows the square root of the average MSE (mean squared error) of 3 confidence interval bounds on the estimated parameter a_i and b_i when the number of participants varies. Observe that the high confidence bounds (i.e., 95% or 90%) fail to bound the root of MSE on a_i or b_i only when the number of participants (M) is very small. This is due to

[1]In principle, there is no incentive for a participant to lie more than 50% of the time, since negating their statements would then give a more accurate truth

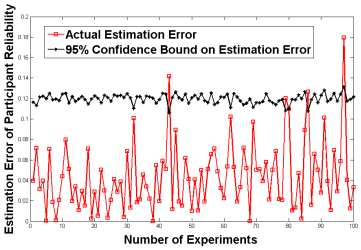
[2]They correspond to one, two and three times standard deviation confidence intervals of normal distribution



(a) 68% Confidence Bound

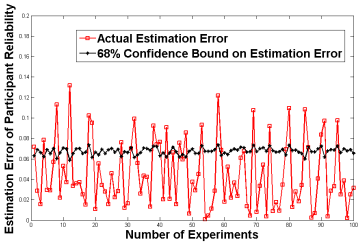


(b) 90% Confidence Bound

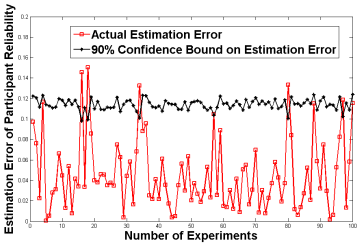


(c) 95% Confidence Bound

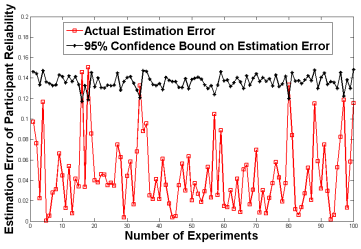
Figure 1: CRB Confidence Bounds on Participant Reliability for Small Observation Matrix



(a) 68% Confidence Bound

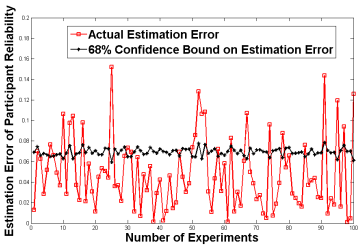


(b) 90% Confidence Bound

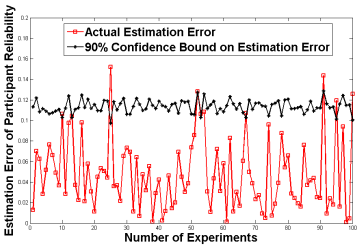


(c) 95% Confidence Bound

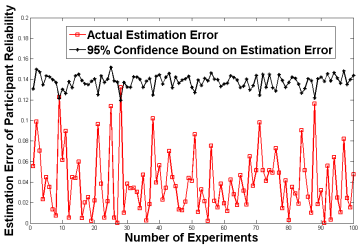
Figure 2: CRB Confidence Bounds on Participant Reliability for Medium Observation Matrix



(a) 68% Confidence Bound

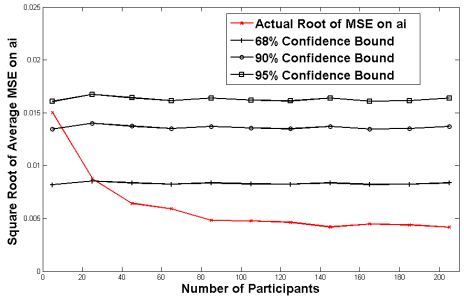


(b) 90% Confidence Bound

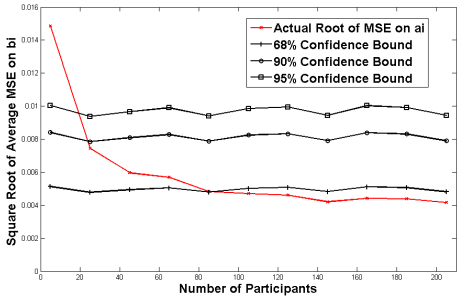


(c) 95% Confidence Bound

Figure 3: CRB Confidence Bounds on Participant Reliability for Large Observation Matrix



(a) Confidence Bound on a_i



(b) Confidence Bound on b_i

Figure 4: CRB Confidence Bound on a_i and b_i versus Vaying M

the poor estimation results of measured varibles when too few participants report their observations. However, when M is reasonablly sufficient (e.g., 25 in the experiment), high confidence bounds always bound the square root of the average MSE on the estimated parameters correctly.

5. RELATED WORK

Social sensing has received significant attention due to the great increase in the number of mobile sensors owned by individuals and the proliferation of Internet connectivity. To assess the credibility of participants and facts reported in participatory sensing and other social sensing applications, a relevant body of work, called *fact-finders*, in the machine learning and data mining communities performs trust analysis. The basic fact-finders include Hubs and Authorities [11],

Observation Matrix Scale	Confidence Level of Estimation	Theoretical Maximum Outliers/Total Experiments	CRB Bound Worst Case Outliers/Total Experiments
Small	68%	32/100	24/100
	90%	10/100	8/100
	95%	5/100	5/100
Medium	68%	32/100	25/100
	90%	10/100	6/100
	95%	5/100	4/100
Large	68%	32/100	25/100
	90%	10/100	9/100
	95%	5/100	2/100

Table 2: CRB Bound on Participant Probability versus Theoretical Results

Average.Log [12], and TruthFinder [15]. Other fact-finders enhance the basic framework by incorporating analysis on properties or dependencies within assertions or sources [9, 2, 8, 7, 6]. Fact-finding in the case of social sensing is more challenging due to the highly dynamic nature of social sensing applications [1]. Moreover, the outputs of fact-finders are generally *rankings* of credibility values of participants and facts. Such rankings cannot be used to directly quantify the participant reliability or fact correctness.

Recent work presented a Bayesian Interpretation scheme [13] representing an initial effort to provide a probability interpretation of ranking outputs from fact-finders. However, it remains an approximation approach in which the accuracy of truth estimation is very sensitive to initial conditions of iterations. Due to this limitation, a maximum likelihood estimation approach using EM algorithm is proposed to provide the first optimal solution to the truth discovery problem in social sensing [14]. The EM scheme was shown to outperform Bayesian interpretation and other state-of-art fact-finders. However, only average estimation accuracies were reported in both of above schemes. The confidence interval of the estimation accuracy has not been found. In contrast, this paper derives, for the first time, the confidence interval of participant reliability based on the approximation of CRB, hence completes the quantification of participant reliability estimation in social sensing.

In estimation theory and statistics, the Cramer-Rao bound refers to a lower bound on the variance of estimators of a deterministic parameter [4]. The bound states the variance of any unbiased estimator is lower-bounded by the inverse of Fisher information [10]. One of the key properties of maximum likelihood estimation is asymptotic normality. An EM scheme provides maximum likelihood estimation of participant reliability for social sensing applications. This paper provides the first quantification approach to compute the confidence interval for participant reliability maximum-likelihood estimation based on the approximation of CRB by leveraging results from estimation and information theory.

6. CONCLUSION

This paper described a quantification approach to compute the confidence interval of the maximum likelihood estimation on participant reliability based on the approximation of CRB in social sensing applications. This quantification approach completely characterizes the estimation performance of participant reliability without knowing the trustworthiness of participants *a priori*. The derived confidence interval is obtained by leveraging the asymptotic normality of maximum likelihood estimation and can be easily computed from the approximated Fisher information contained in participants' reliability estimations. Evaluation results show that the error in the estimated participant reliability is well bounded by the computed bound. In future work, a tighter CRB bound can probably be derived without knowing the truthfulness of each measured variable. This actual CRB bound is expected to better track the actual MSE of the estimated parameters.

Acknowledgements

Research reported in this paper was sponsored by the Army Research Laboratory 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.

7. 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] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE 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. Dong, L. Berti-Equille, and D. Srivastava. Truth discovery and copying detection in a dynamic world. *VLDB*, 2(1):562–573, 2009.

[8] 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.

[9] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *WSDM*, pages 131–140, 2010.

[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. *Journal of the ACM*, 46(5):604–632, 1999.

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

[13] 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.

[14] D. Wang, T. Abdelzaher, and L. Kaplan. On truth discovery in social sensing: A maximum likelihood estimation approach. UIUC Technical Report, http://hdl.handle.net/2142/25815, 2011.

[15] 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.