# On Truth Discovery in Social Sensing: A Maximum Likelihood Estimation Approach

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

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

***Abstract*—This paper addresses the challenge of truth discovery from noisy social sensing data. The work is motivated by the emergence of social sensing as a data collection paradigm of growing interest, where humans perform sensory data collection tasks. A challenge in social sensing applications lies in the noisy nature of data. Unlike the case with well-calibrated and well-tested infrastructure sensors, humans are less reliable, and the likelihood that participants' measurements are correct is often unknown *a priori*. Given a set of human participants of unknown trustworthiness together with their sensory measurements, this paper poses the question of whether one can use this information alone to determine, in an analytically founded manner, the probability that a given measurement is true. The paper focuses on binary measurements. While some previous work approached the answer in a heuristic manner, we offer the first *optimal solution* to the above truth discovery problem. Optimality, in the sense of maximum likelihood estimation, is attained by solving an expectation maximization problem that returns the best guess regarding the correctness of each measurement. The approach is shown to outperform the state of the art fact-finding heuristics, as well as simple baselines such as majority voting.**

***Keywords*-truth discovery, social sensing, maximum likelihood estimation, expectation maximization**

## I. Introduction

This paper presents a maximum likelihood estimation approach to truth discovery from social sensing data. Social sensing has emerged as a new paradigm for collecting sensory measurements by means of "crowd-sourcing" sensory data collection tasks to a human population. The paradigm is made possible by the proliferation of a variety of sensors in the possession of common individuals, together with networking capabilities that enable data sharing. Examples includes cell-phone accelerometers, cameras, GPS devices, smart power meters, and interactive game consoles (e.g., Wii). Individuals who own such sensors can thus engage in data collection for some purpose of mutual interest. A classical example is geotagging campaigns, where participants report locations of conditions in their environment that need attention (e.g., litter in public parks).

A significant challenge in social sensing applications lies in ascertaining the correctness of collected data. Data collection is often open to a large population. Hence, the participants and their reliability are typically not known *a priori*. The term, participant (or source) *reliability* is used in this paper to denote the probability that the participant reports correct observations. Reliability may be impaired because of poor used sensor quality, lack of sensor calibration, lack of (human) attention to the task, or even intent to deceive. The question posed in this paper is whether or not we can determine, given only the measurements sent and without knowing the reliability of sources, which of the reported observations are true and which are not. In this paper, we concern ourselves with (arrays of) binary measurements only; for example, reporting whether or not litter exists at each of multiple locations of interest. We develop a maximum likelihood estimator that assigns truth values to measurements without prior knowledge of source reliability. The algorithm makes inferences regarding both source reliability and measurement correctness by observing which observations coincide and which don't. It is shown to be surprisingly accurate in assessing measurement correctness as long as sources, on average, make multiple observations, and as long as some sources make the same observation.

Note that, a trivial way of accomplishing the truth discovery task is by "believing" only those observations that are reported by a sufficient number of sources. We call such a scheme, *voting*. The problem with voting schemes is that they do not attempt to infer source reliability and do not take that estimate into account. Hence, observations made by several unreliable sources may be believed over those made by a few reliable ones. Instead, we cast the truth discovery problem as one of joint maximum likelihood estimation of both source reliability and observation correctness. We solve the problem using the Expectation maximization (EM) algorithm.

Expectation maximization (EM) is a general optimization technique for finding the maximum likelihood estimation of parameters in a statistic model where the data are "incomplete" [9]. It iterates between two main steps (namely, the E-step and the M-step) until the estimation converges (i.e., the likelihood function reaches the maximum). The paper shows that social sensing applications lend themselves nicely to an EM formulation. The optimal solution, in the sense of maximum likelihood estimation, directly leads to an

accurate quantification of measurement correctness as well as participant reliability. Moreover, the solution is shown to be simple and easy to implement.

Finally, one should observe that the truth discovery problem, as formulated in this paper, is not an invention of the authors. Prior literature attempted to solve it using heuristics whose inspiration can be traced back to Google's PageRank [6]. PageRank iteratively ranks the credibility of sources on the Web, by iteratively considering the credibility of sources who link to them. Extensions of PageRank, known as fact-finders, iteratively compute the credibility of sources and claims. Specifically, they estimate the credibility of claims from the credibility of sources that make them, then estimate the credibility of sources based on the credibility of their claims. Several algorithms exist that feature modifications of the above basic heuristic scheme [5], [13], [20], [26], [27]. In contrast, ours is the first attempt to optimally solve the problem by casting it as one of expectation maximization.

We evaluate our algorithm in simulation, as well as using an emulated geotagging application scenario. Evaluation results show that the proposed maximum likelihood scheme outperforms the state-of-art heuristics as well as simple baselines (voting) in quantifying the probability of measurement correctness and participant reliability.

The rest of this paper is organized as follows: In Section II, we present the fact-finding model for social sensing applications. The proposed maximum likelihood estimation approach is discussed in Section III. Implementation and evaluation results are presented in Section IV. We review related work in Section V. Finally, we conclude the paper in Section VI.

## II. The Problem Formulation of Social Sensing

To formulate the truth discovery problem in social sensing in a manner amenable to rigorous optimization, 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. For example, a group of individuals interested in the appearance of their neighborhood might join a sensing campaign to report all locations of offensive graffiti. Alternatively, a group of drivers might join a campaign to report freeway locations in need of repair. Hence, each measured variable denotes the existence or lack thereof of an offending condition at a given location[1]. In this effort, we consider only binary variables and assume, without loss of generality, that their "normal" state is negative (e.g., no offending graffiti on walls, or no potholes on streets). Hence, participants report only when a positive value is encountered.

Each participant generally observes only a subset of all variables (e.g., the conditions at locations they have been to). Our goal is to determine which observations are correct and which are not. As mentioned in the introduction, we differ from a large volume of previous sensing literature in that we assume no prior knowledge of source reliability, as well as no prior knowledge of the correctness of individual observations.

Let $S_iC_j$ denote an observation reported by participant $S_i$ claiming that $C_j$ is true (e.g., that graffiti is found at a given location, or that a given street is in disrepair). 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 input to our 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. For example, it may represent the probability that any street, in general, is in disrepair. Note that, this value can be known from past statistics. It does not indicate, however, whether any particular claim about disrepair at a particular location 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:

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

[1]We assume that locations are discretized, and therefore finite. For example, they are given by street addresses or mile markers.

The goal of the algorithm is to compute (i) the best estimate $h_j$ of the value each variable $C_j$ and (ii) the best estimate $e_i$ of the reliability of each participant $S_i$. Let us denote the sets of the estimates by vectors $H$ and $E$, respectively. Our goal is to find the optimal $H^*$ and $E^*$ vectors in the sense of being most consistent with the observation matrix $SC$. Formally, this is given by:

$$< H^*, E^* >= \underset{<H,E>}{\operatorname{argmax}}\, p(SC|H,E) \tag{5}$$

## III. Expectation Maximization

We solve the problem formulated in the previous section using the Expectation-Maximization (EM) algorithm. It is a general algorithm for finding the maximum likelihood estimates of parameters in a statistic model, where the data are "incomplete" or the likelihood function involves latent variables [9]. Intuitively, what EM does is iteratively "completes" the data by "guessing" the values of hidden variables then re-estimates the parameters by using the guessed values as true values.

### A. Mathematical Formulation

Much like finding a Lyapunov function to prove stability, the main challenge in using the EM algorithm lies in the mathematical formulation of the problem in a way that is amenable to an EM solution. Given an observed data set $X$, one should judiciously choose the set of latent or missing values $Z$, and a vector of unknown parameters $\theta$, then formulate a likelihood function $L(\theta; X, Z) = p(X, Z|\theta)$, such that the maximum likelihood estimate (MLE) of the unknown parameters $\theta$ is decided by:

$$L(\theta; X) = p(X|\theta) = \sum_{Z} p(X, Z|\theta) \tag{6}$$

Once the formulation is complete, the EM algorithm finds the maximum likelihood estimate by iteratively performing the following steps:

- E-step: Compute the expected log likelihood function where the expectation is taken with respect to the computed conditional distribution of the latent variables given the current settings and observed data.

$$Q\left(\theta|\theta^{(t)}\right) = E_{Z|X,\theta^{(t)}}[\log L(\theta; X, Z)] \tag{7}$$

- M-step: Find the parameters that maximize the $Q$ function in the E-step to be used as the estimate of $\theta$ for the next iteration.

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q\left(\theta|\theta^{(t)}\right) \tag{8}$$

Our participatory sensing problem fits nicely into the Expectation Maximization (EM) model. First, we introduce a latent variable $Z$ for each measured variable to indicate whether it is true or not. Specifically, we have a corresponding variable $z_j$ for the $j^{th}$ measured variable such that: $z_j = 1$ when the measured variable $C_j$ is true and $z_j = 0$ otherwise. We further denote the observation matrix $SC$ as the observed data $X$, and take $\theta = (a_1, a_2, ...a_M; b_1, b_2, ...b_M)$ as the parameter of the model that we want to estimate. The goal is to get the maximum likelihood estimate of $\theta$ for the model containing observed data $X$ and latent variables $Z$.

The likelihood function $L(\theta; X, Z)$ is given by:

$$\begin{aligned} &L(\theta; X, Z) = p(X, Z|\theta) \\ &= \prod_{j=1}^{N} \left\{ \prod_{i=1}^{M} a_i^{S_i C_j} (1 - a_i)^{(1 - S_i C_j)} \times d \times z_j \right. \\ &\left. + \prod_{i=1}^{M} b_i^{S_i C_j} (1 - b_i)^{(1 - S_i C_j)} \times (1 - d) \times (1 - z_j) \right\} \end{aligned} \tag{9}$$

### B. Deriving the E-Step and M-Step

Given the above formulation, the Expectation step (E-step) becomes:

$$\begin{aligned} &Q\left(\theta|\theta^{(t)}\right) = E_{Z|X,\theta^{(t)}}[\log L(\theta; X, Z)] \\ &= \sum_{j=1}^{N} \left\{ p(z_j = 1|X_j, \theta^{(t)}) \right. \\ &\times \left[ \sum_{i=1}^{M} (S_i C_j \log a_i + (1 - S_i C_j) \log(1 - a_i)) + \log d \right] \\ &+ p(z_j = 0|X_j, \theta^{(t)}) \\ &\left. \times \left[ \sum_{i=1}^{M} (S_i C_j \log b_i + (1 - S_i C_j) \log(1 - b_i) + \log(1 - d)) \right] \right\} \end{aligned} \tag{10}$$

where $X_j$ is the $j^{th}$ column of the observed $SC$ matrix and $p(z_j = 1|X_j, \theta^{(t)})$ is the conditional probability of the latent variable $z_j$ to be true given the observation matrix related to the $j^{th}$ measured variable and current estimate of $\theta$, which is given by:

$$\begin{aligned} &p(z_j = 1|X_j, \theta^{(t)}) \\ &= \frac{p(z_j = 1; X_j, \theta^{(t)})}{p(X_j, \theta^{(t)})} \\ &= \frac{p(X_j, \theta^{(t)}|z_j = 1)p(z_j = 1)}{p(X_j, \theta^{(t)}|z_j = 1)p(z_j = 1) + p(X, \theta^{(t)}|z_j = 0)p(z_j = 0)} \\ &= \frac{A(t, j) \times d}{A(t, j) \times d + B(t, j) \times (1 - d)} \\ &= Z(t, j) \end{aligned} \tag{11}$$

where $A(t,j)$ and $B(t,j)$ are defined as:

$$\begin{aligned} A(t,j) &= p(X_j, \theta^{(t)} | z_j = 1) \\ &= \prod_{i=1}^{M} a_i^{(t)S_iC_j} (1 - a_i^{(t)})^{(1-S_iC_j)} \\ B(t,j) &= p(X_j, \theta^{(t)} | z_j = 0) \\ &= \prod_{i=1}^{M} b_i^{(t)S_iC_j} (1 - b_i^{(t)})^{(1-S_iC_j)} \end{aligned} \tag{12}$$

Next we simplify Equation (10) by noting that the conditional probability of $p(z_j = 1 | X_j, \theta^{(t)})$ is only a function of $t$ and $j$. Thus, we represent it by $Z(t,j)$. Similarly, $p(z_j = 0 | X, \theta^{(t)})$ is simply:

$$\begin{aligned} & p(z_j = 0 | X, \theta^{(t)}) \\ &= 1 - p(z_j = 1 | X, \theta^{(t)}) \\ &= \frac{B(t,j) \times (1-d)}{A(t,j) \times d + B(t,j) \times (1-d)} \\ &= 1 - Z(t,j) \end{aligned} \tag{13}$$

Substituting from Equation (11) and (13) into Equation (10), we get:

$$\begin{aligned} & Q\left(\theta | \theta^{(t)}\right) \\ &= \sum_{j=1}^{N} \Bigg\{ Z(t,j) \\ &\times \left[ \sum_{i=1}^{M} (S_iC_j \log a_i + (1 - S_iC_j) \log(1 - a_i)) + \log d \right] \\ &+ (1 - Z(t,j)) \\ &\times \left[ \sum_{i=1}^{M} (S_iC_j \log b_i + (1 - S_iC_j) \log(1 - b_i)) + \log(1-d) \right] \Bigg\} \end{aligned} \tag{14}$$

The Maximization step (M-Step) is given by Equation (8). We choose $\theta^*$ (i.e., $(a_1^*, a_2^*, ...a_M^*; b_1^*, b_2^*, ...b_M^*)$) that maximizes the $Q\left(\theta | \theta^{(t)}\right)$ function in each iteration to be the $\theta^{(t+1)}$ of the next iteration.

To get $\theta^*$ that maximizes $Q\left(\theta | \theta^{(t)}\right)$, we set the derivatives $\frac{\partial Q}{\partial a_i} = 0$, $\frac{\partial Q}{\partial b_i} = 0$, which yields:

$$\begin{aligned} & \sum_{j=1}^{N} \left[ Z(t,j)(S_iC_j \frac{1}{a_i^*} - (1 - S_iC_j) \frac{1}{1 - a_i^*}) \right] = 0 \\ & \sum_{j=1}^{N} \left[ (1 - Z(t,j))(S_iC_j \frac{1}{b_i^*} - (1 - S_iC_j) \frac{1}{1 - b_i^*}) \right] = 0 \end{aligned} \tag{15}$$

Let us define $SJ_i$ as the set of measured variables the participant $S_i$ actually observes in the observation matrix (i.e, $SC$), and $\bar{SJ}_i$ as the set of measured variables participant $S_i$ does not observe in the observation matrix. Thus, Equation (15) can be rewritten as:

$$\begin{aligned} & \sum_{j \in SJ_i} Z(t,j) \frac{1}{a_i^*} - \sum_{j \in \bar{SJ}_i} Z(t,j) \frac{1}{1 - a_i^*} = 0 \\ & \sum_{j \in SJ_i} (1 - Z(t,j)) \frac{1}{b_i^*} - \sum_{j \in \bar{SJ}_i} (1 - Z(t,j)) \frac{1}{1 - b_i^*} = 0 \end{aligned} \tag{16}$$

Solving the above equations, we can get expressions of the optimal $a_i^*$ and $b_i^*$:

$$\begin{aligned} a_i^{(t+1)} = a_i^* &= \frac{\sum_{j \in SJ_i} Z(t,j)}{\sum_{j=1}^{N} Z(t,j)} \\ b_i^{(t+1)} = b_i^* &= \frac{K_i - \sum_{j \in SJ_i} Z(t,j)}{N - \sum_{j=1}^{N} Z(t,j)} \end{aligned} \tag{17}$$

where $K_i$ is the number of measured variables observed by participant $S_i$ and $N$ is the total number of measured variables in the observation matrix. $Z(t,j)$ is defined in Equation (11).

Given the above, The E-step and M-step of EM optimization reduce to simply calculating Equation (11) and Equation (17) iteratively until they converge. Since the measured variable is binary, we can compute the optimal decision vector $H^*$ from the converged value of $Z(t,j)$. Specially, $h_j$ is true if $Z(t,j) \geq 0.5$ and false otherwise. At the same time, we can also compute the optimal estimation vector $E^*$ from the converged values of either $a_i^{(t)}$ or $b_i^{(t)}$ based on Equation (4). This completes the mathematical development. We summarize the resulting algorithm in the subsection below.

### C. The Final Algorithm

Given the observation matrix $SC$, our algorithm begins by initializing the parameter $\theta$ with random values between 0 and 1[2]. The algorithm then performs the E-steps and M-steps iteratively until $\theta$ converges. Specifically, we compute the conditional probability of a measured variable to be true (i.e., $Z(t,j)$) from Equation (11) and the probability that a participant observes a measured variable given the variable is true or false (i.e., $a_i^{(t+1)}, b_i^{(t+1)}$) from Equation (17). After the estimated value of $\theta$ converges, we compute the optimal decision vector $H^*$ (i.e., decide whether each measured variable $C_j$ is true or not) based on the converged value of $Z(t,j)$ (i.e., $Z_j^c$). We can also compute the optimal estimation vector $E^*$ (i.e., the estimated $t_i$ of each participant) from the converged values of $a_i^{(t)}$ or $b_i^{(t)}$ (i.e., $a_i^c$ or $b_i^c$) based on Equation (4) as shown in the pseudocode in Algorithm 1.

[2]In practice, if the a rough estimate of the average reliability of participants is known *a priori*, EM will converge faster

**Algorithm 1** Expectation Maximization Algorithm

Initialize $\theta$ with random values between 0 and 1
**while** $\theta^{(t)}$ does not converge **do**
  **for** $j = 1 : N$ **do**
    compute $Z(t, j)$ based on Equation (11)
  **end for**
  $\theta^{(t+1)} = \theta^{(t)}$
  **for** $i = 1 : M$ **do**
    compute $a_i^{(t+1)}, b_i^{(t+1)}$ based on Equation (17)
    update $a_i^{(t)}, b_i^{(t)}$ with $a_i^{(t+1)}, b_i^{(t+1)}$ in $\theta^{(t+1)}$
  **end for**
  $t = t + 1$
**end while**
Let $Z_j^c$ = converged value of $Z(t, j)$
Let $a_i^c$ = converged value of $a_i^{(t)}$; $b_i^c$ = converged value of $b_i^{(t)}$
**for** $j = 1 : N$ **do**
  **if** $Z_j^c \geq 0.5$ **then**
    $h_j^*$ is true
  **else**
    $h_j^*$ is false
  **end if**
**end for**
**for** $i = 1 : M$ **do**
  calculate $e_i^*$ from $a_i^c$ or $b_i^c$ based on Equation (4)
**end for**
Return the computed optimal estimates of measured variables $C_j = h_j^*$ and source reliability $e_i^*$.

## IV. Evaluation

In this section, we carry out simulation experiments to evaluate the performance of the proposed EM scheme in terms of estimation accuracy of the probability that a participant is right or a measured variable is true compared to other state-of-art solutions. We begin by considering algorithm performance for different abstract observation matrices (SC), then apply it to a simulated participatory sensing application. We show that the new algorithm outperforms the state of the art.

### A. A Simulation Study

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). Remember that, as stated in our application model, 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[3].

In recent work, the authors demonstrated a heuristic, called *Bayesian Interpretation* [24], that outperformed all contenders from prior literature. We first compare our optimal EM scheme against the Bayesian Interpretation heuristic, under simulation conditions reported in the previous paper. We then consider more challenging conditions not investigated in [24], and compare EM to four state of the art algorithms (including Bayesian Interpretation) in that scenario. Results show a significant performance improvement over all heuristics compared.

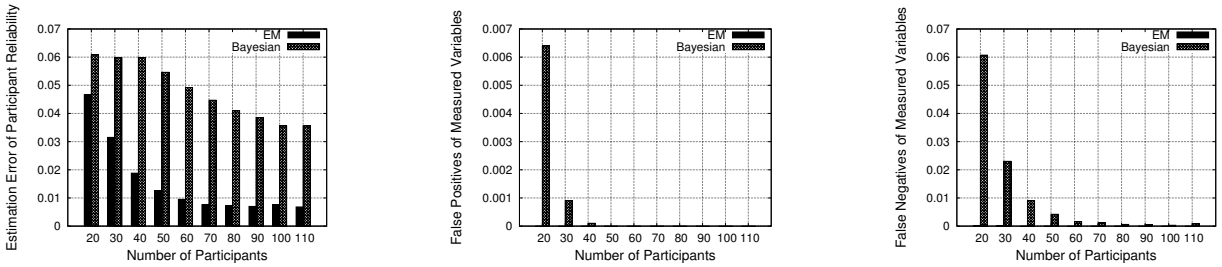
In the first scenario, we compare the estimation accuracy of EM and the Bayesian Interpretation scheme by varying the number of participants in the system. The number of reported measured variables was fixed at 2000, of which 1000 variables were reported correctly and 1000 were misreported. The average number of observations per participant was set to 100. The number of participants was varied from 20 to 110. Results (computed probability estimates) were averaged over 100 experiments involving the same sources and variables. Figure 1 compares the accuracy of the two schemes. Observe that EM has a much lower estimation error in participant reliability (i.e., the probability that a participant is right) compared to the Bayesian Interpretation scheme, and zero false positives and false negatives (i.e., misclassified observations). False positives denote variables misclassified as true (e.g., locations determined to contain potholes where in fact they do not). False negatives denote variables misclassified as false (i.e., real pothole reports classified as incorrect).

The second scenario compares the two schemes when the average number of observations per participant changes. As before, we fix the number of correctly and incorrectly reported variables to 1000 respectively. We also set the number of participants to 30. The average number of observations per participant is varied from 100 to 1000. Results are averaged over 100 experiments. Figure 2 shows these results. Note that, the participant reliability estimation error of the EM scheme remains at a much lower level compared to the Bayesian Interpretation scheme. No false positives or false negatives are observed.

The third experiment examines the effect of changing the measured variable mix on the estimation accuracy of two schemes. We vary the ratio of the number of correctly reported variables to the total number of reported variables from 0.1 to 0.6, while fixing the total number of such variables to 2000. The number of participants is fixed at 30 and the average number of observations per participant is set to 150. Results are averaged over 100 experiments. These results are shown in Figure 3. Again, we observe that EM scheme continues to outperform the Bayesian Interpretation scheme in both participant reliability and measured variable estimation.

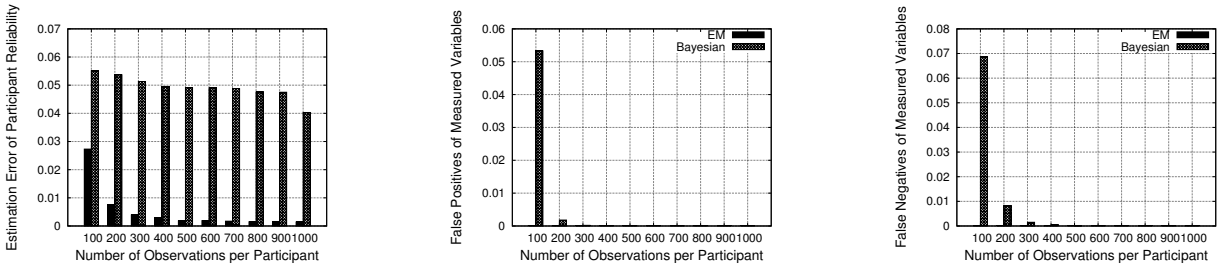
As done in [24], the performance comparison between EM and Bayesian Interpretation averages results over multiple observation matrices. This is intended to approximate performance where multiple matrices are reported (e.g., from

[3]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



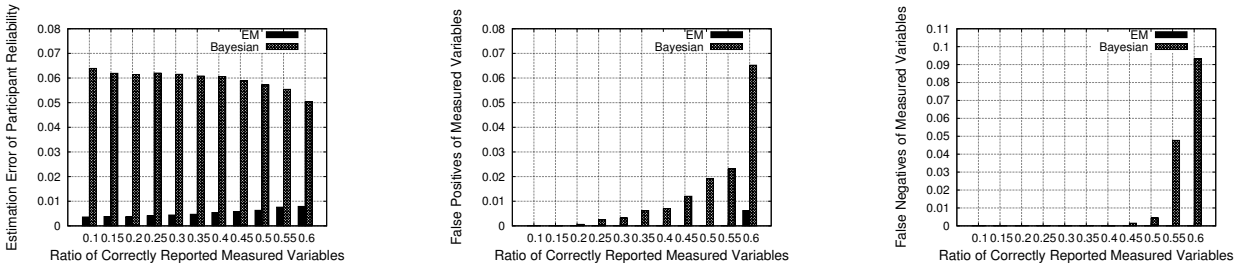

(a) Participant Reliability Estimation Accuracy (b) Measured Variable Estimation: False Positives (c) Measured Variable Estimation: False Negatives

Figure 1. Estimation Accuracy versus Number of Participants in Dense Sensing




(a) Participant Reliability Estimation Accuracy (b) Measured Variable Estimation: False Positives (c) Measured Variable Estimation: False Negatives

Figure 2. Estimation Accuracy versus Average Number of Observations per Participant in Dense Sensing



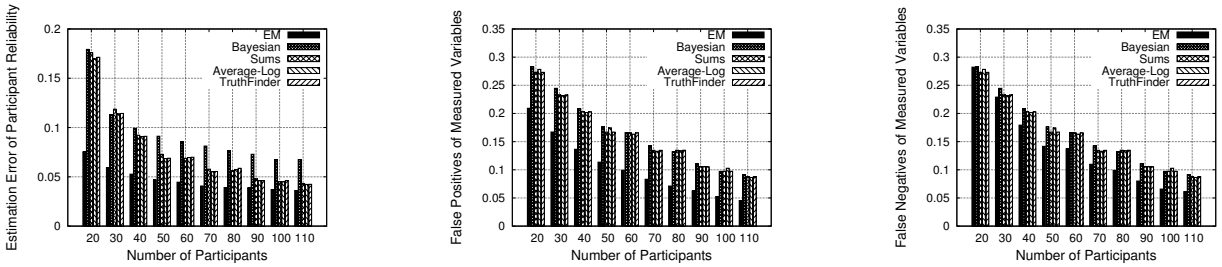

(a) Participant Reliability Estimation Accuracy (b) Measured Variable Estimation: False Positives (c) Measured Variable Estimation: False Negatives

Figure 3. Estimation Accuracy versus Ratio of Correctly Reported Measured Variables in Dense Sensing

successive observation intervals) involving the same set of sources and measured variables. The next question to answer is: will results still be accurate if only one observation matrix is available?

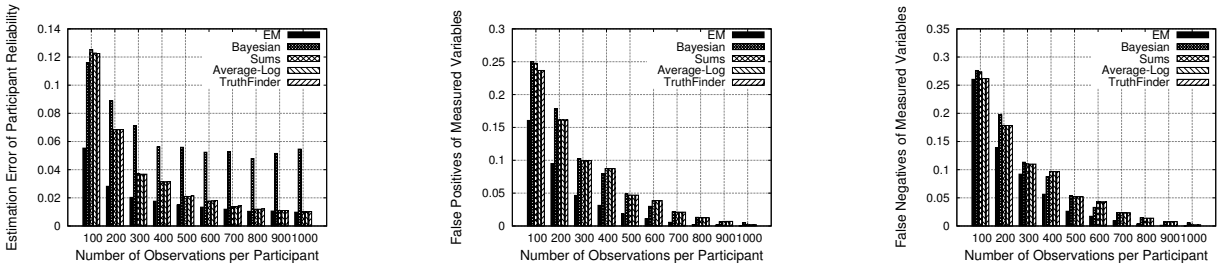
The above question is answered by comparing EM, Bayesian Interpretation and three previous fact-finder schemes from prior literature that can function using only the inputs offered in our problem formulation. These fact-finders are Sums [18], Average-Log [20], and TruthFinder [26]. We repeated the above experiments now using a single observation matrix. Reported results are averaged over 100 random participant correctness probability distributions. First, we show the performance comparison between EM and other schemes by varying the number of participants in the network. Results are shown in Figure 4. Observe that EM has the smallest estimation error on participant reliability and the least false positives among all schemes under comparison. For false negatives, EM performs similarly to other schemes when the number of participants is small and starts to gain improvements when the number of participants becomes large. Note also that the performance gain of EM becomes large when the number of participants is small, illustrating that EM is more useful when the observation matrix is sparse.

We then repeat the second experiment to compare the EM scheme with other baselines while varying the average number of observations per participant. The results are shown in Figure 5. Observe that EM outperforms all baselines in terms of both participant reliability estimation accuracy and false positives as the average number of observations per participant changes. For false negatives, EM has similar performance as other baselines when the average



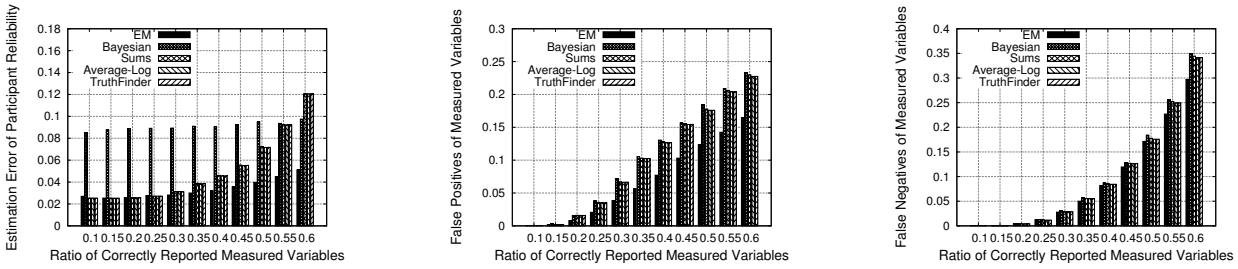

(a) Participant Reliability Estimation Accuracy (b) Measured Variable Estimation: False Positives (c) Measured Variable Estimation: False Negatives

Figure 4. Estimation Accuracy versus Number of Participants in Sparse Sensing




(a) Participant Reliability Estimation Accuracy (b) Measured Variable Estimation: False Positives (c) Measured Variable Estimation: False Negatives

Figure 5. Estimation Accuracy versus Average Number of Observations per Participant in Sparse Sensing




(a) Participant Reliability Estimation Accuracy (b) Measured Variable Estimation: False Positives (c) Measured Variable Estimation: False Negatives

Figure 6. Estimation Accuracy versus Ratio of Correctly Reported Measured Variables in Sparse Sensing

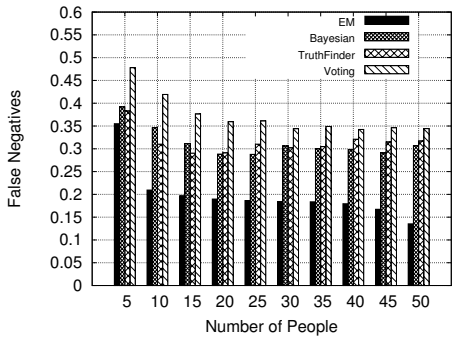
number of observations per participant is small and starts to gain advantage as the average number of observations per participant becomes large. As before, the performance gain of EM is higher when the average number of observations per participant is low, verifying once more the high accuracy of EM for sparser observation matrices.

Finally, we repeated the third experiment comparing the EM scheme to other baselines while varying the ratio of the number of correctly reported variables to the total number of measured variables in the network. We observe that EM has almost the same performance as other fact-finder baselines when the fraction of correctly reported variables is relatively small. However, as the number of variables (correctly) reported as true grows, EM is shown to have a better performance in both participant reliability and measured variable estimation. This concludes our general simulations. In the next section, we simulate the performance of a specific social sensing application.

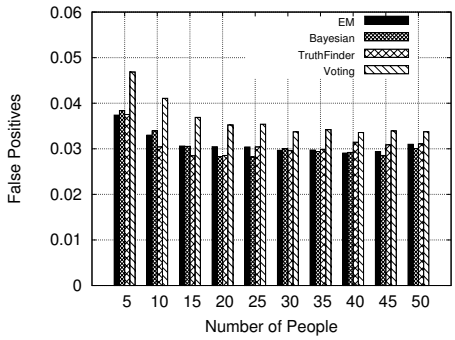
### *B. A Geotagging Case Study*

In this section, we applied the proposed EM scheme to a typical participatory sensing application: Geotagging locations of litter in a park or hiking area. In this application, litter may be found along the trails (usually proportionally to their popularity). Participants visiting the park geotag and report locations of litter. Their reports are not reliable however, erring both by missing some locations, as well as misrepresenting other objects as litter. The goal of the application is to find where litter is actually located in the park, while disregarding all false reports.

To evaluate the performance of different schemes, we define two metrics of interest: (i) *false negatives* defined

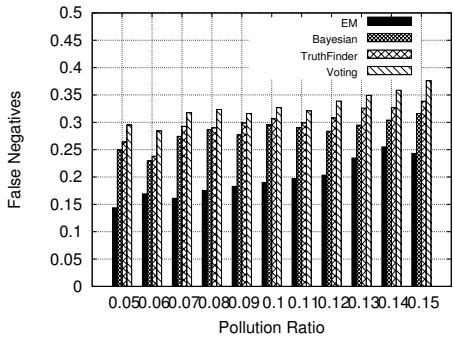



(a) False Negatives (missed litter/total litter)

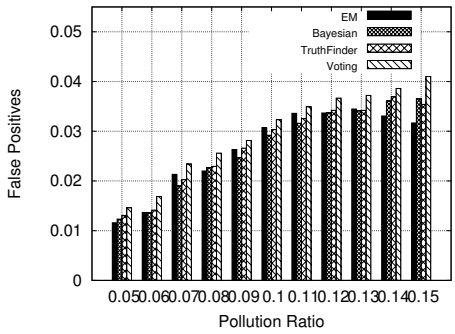



(b) False Positives (false locations/total locations)

Figure 7. Litter Geotagging Accuracy versus Number of People




(a) False Negatives (missed litter/total litter)




(b) False Positives (false locations/total locations)

Figure 8. Litter Geotagging Accuracy versus Pollution Ratio of Park

as the ratio of litter locations missed by the EM scheme to the total number of litter locations in the park, and (ii) *false positives* defined as the ratio of the number of incorrectly labeled locations by EM, to the total number of locations in the park. We compared the proposed EM scheme to the Bayesian Interpretation scheme and to voting, where locations are simply ranked by the number of times people report them.

We created a simplified trail map of a park, represented by a binary tree. The entrance of the park (e.g., where parking areas are usually located) is the root of the tree. Internal nodes of the tree represent forking of different trails. We assume trails are quantized into discretely labeled locations (e.g., numbered distance markers). In our simulation, at each forking location along the trails, participants have a certain probability $P_c$ to continue walking and $1 - P_c$ to stop and return. Participants who decide to continue have equal probability to select the left or right path. The majority of participants are assumed to be reliable (i.e., when they geotag and report litter at a location, it is more likely than not that the litter exists at that location).

In the first experiment, we study the effect of the number of people visiting the park on the estimation accuracy of different schemes. We choose a binary tree with a depth of 4 as the trail map of the park. Each segment of the trail (between two forking points) is quantized into 100 potential locations (leading to 1500 discrete locations in total on all trails). We define the pollution ratio of the park to be the ratio of the number of littered locations to the total number of locations in the park. The pollution ratio is fixed at 0.1 for the first experiment. The probability that people continue to walk past a fork in the path is set to be 95% and the percent of reliable participants is set to be 80%. We vary the number of participants visiting the park from 5 to 50. The corresponding estimation results of different schemes are shown in Figure 7. Observe that both false negatives and false positives decrease as the number of participants increases for all schemes. This is intuitive: the chances of finding litter on different trails increase as the number of people visiting the park increases. Note that, the EM scheme outperforms others in terms of false negatives, which means EM can find more pieces of litter than other schemes under

the same conditions. The improvement becomes significant (i.e., around 20%) when there is a sufficient number of people visiting the park. For the false positives, EM performs similarly to Bayesian Interpretation and Truth Finder scheme and better than voting. Generally, voting performs the worst in accuracy because it simply counts the number of reports complaining about each location but ignores the reliability of individuals who make them.

In the second experiment, we show the effect of park pollution ratio (i.e, how littered the park is) on the estimation accuracy of different schemes. The number of individuals visiting the park is set to be 40. We vary the pollution ratio of the park from 0.05 to 0.15. The estimation results of different schemes are shown in Figure 8. Observe that both the false negatives and false positives of all schemes increase as the pollution ratio increases. The reason is that: litter is more frequently found and reported at trails that are near the entrance point. The amount of unreported litter at trails that are far from entrance increases more rapidly compared to the total amount of litter as the pollution ratio increases. Note that, the EM scheme continues to find more actual litter compared to other baselines. The performance of false positives is similar to other schemes.

The evaluation demonstrates that the new EM scheme generally outperforms the current state of the art in inferring facts from participatory sensing data.

## V. Related Work

Social sensing, and participatory sensing in particular has received significant attention due to the great increase in the number of mobile sensors owned by individuals (e.g., smart phones with GPS, camera and etc.) and the proliferation of Internet connectivity to upload and share sensed data (e.g., WiFi and 4G networks). The concept of participatory sensing was first introduced in [7]. A broad overview of such applications is presented in [1]. Some early applications of participatory sensing include CenWits [15], a participatory sensor network to search and rescue hikers in emergency situations, CarTel [17], a vehicular sensor network for traffic monitoring and mitigation, and BikeNet [12], a bikers sensor network for sharing cycling related data and mapping the cyclist experience. More recent work has focused on addressing the challenges of preserving privacy and building general models in sparse and multi-dimensional social sensing space [2], [3], [14]. Participatory sensing is often organized as "sensing campaigns" where participants are recruited to contribute their personal measurements as part of a large-scale effort to collect data about a population or a geographical area. Examples include documenting the quality of roads [22], the level of pollution in a city [19], or reporting garbage cans on campus [21]. In addition to sensing applications where participants are recruited, social sensing can also be triggered spontaneously without prior coordination (e.g., via Twitter and Youtube). Recent research attempts to understand the fundamental factors that affect the behavior of these emerging social sensing applications, such as analysis of characteristics of social networks [8], information propagation [16] and tipping points [25].

To assess the credibility of facts reported in participatory sensing and other social sensing applications, a relevant body of work in the machine learning and data mining communities performs trust analysis. Hubs and Authorities [18], for example, used a basic fact-finder where the belief in an assertion $c$ is $B(c) = \sum_{s \in S_c} T(s)$ and the truthfulness of a source $s$ is $T(s) = \sum_{c \in C_s} B(c)$, where $S_c$ and $C_s$ are the sources claiming a given assertion and the assertions claimed by a particular source, respectively. Pasternack *et al.* extend the fact-finder framework by incorporating prior knowledge into the analysis and proposes several extended algorithms: *Average.Log, Investment, Pooled Investment* [20]. Yin et al introduce *TruthFinder* as an unsupervised fact-finder for trust analysis on a providers-facts network [26]. Other fact-finders enhance the basic framework by incorporating analysis on properties or dependencies within assertions or sources. Galland et al. [13] take the notion of hardness of facts into consideration by proposing their algorithms: *Cosine, 2-Estimates, 3-Estimates*. The source dependency detection problem has been discussed and several solutions have been proposed [5], [10], [11]. Additionally, trust analysis has been done both on a homogeneous network [4], [27] and a heterogeneous network [23]. Our proposed EM scheme is the first piece of work that finds a maximum likelihood estimator to directly and optimally quantify the accuracy of conclusions obtained from credibility analysis in social sensing.

The Bayesian Interpretation scheme [24] comes closest to our current work. It represents an initial effort to lay out solid analytical foundations for fact-finding. However, the Bayesian Interpretation remains an approximation approach in which the accuracy of truth estimation is very sensitive to initial conditions of iterations. Due to this limitation, as shown in the current paper, EM outperforms Bayesian Interpretation.

In this paper, we intentionally start with a simplified application model, where the measured variables are binary, measurements are independent, and participants do not influence each other's reports (e.g., do not propagate each other's rumors). Subsequent work will address the above limitations.

## VI. Conclusion

This paper described a maximum likelihood estimation approach to accurately discover the truth in social sensing applications. The approach can determine the correctness of reported observations given only the measurements sent without knowing the trustworthiness of participants. The optimal solution is obtained by solving an expectation maximization problem and can directly lead to an analytically founded quantification of the correctness of measurements

as well as the reliability of participants. Evaluation results show that non-trivial estimation accuracy improvements can be achieved by the proposed maximum likelihood estimation approach compared to other state of the art solutions.

## REFERENCES


[1] T. Abdelzaher et al. Mobiscopes for human spaces. *IEEE Pervasive Computing*, 6(2):20–29, 2007.

[2] H. Ahmadi, T. Abdelzaher, J. Han, N. Pham, and R. Ganti. The sparse regression cube: A reliable modeling technique for open cyber-physical systems. In *Proc. 2nd International Conference on Cyber-Physical Systems (ICCPS'11)*, 2011.

[3] H. Ahmadi, N. Pham, R. Ganti, T. Abdelzaher, S. Nath, and J. Han. Privacy-aware regression modeling of participatory sensing data. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, SenSys '10, pages 99–112, New York, NY, USA, 2010. ACM.

[4] R. Balakrishnan. Source rank: Relevance and trust assessment for deep web sources based on inter-source agreement. In *20th World Wide Web Conference (WWW'11)*, 2011.

[5] 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.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *7th international conference on World Wide Web (WWW'07)*, pages 107–117, 1998.

[7] J. Burke et al. Participatory sensing. In *Workshop on World-Sensor-Web (WSW): Mobile Device Centric Sensor Networks and Applications*, pages 117–134, 2006.

[8] S. A. Delre, W. Jager, and M. A. Janssen. Diffusion dynamics in small-world networks with heterogeneous consumers. *Comput. Math. Organ. Theory*, 13:185–202, June 2007.

[9] 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.

[10] X. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava. Global detection of complex copying relationships between sources. *PVLDB*, 3(1):1358–1369, 2010.

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

[12] S. B. Eisenman et al. The bikenet mobile sensing system for cyclist experience mapping. In *SenSys'07*, November 2007.

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

[14] 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.

[15] J.-H. Huang, S. Amjad, and S. Mishra. CenWits: a sensor-based loosely coupled search and rescue system using witnesses. In *SenSys'05*, pages 180–191, 2005.

[16] C. Hui, M. K. Goldberg, M. Magdon-Ismail, and W. A. Wallace. Simulating the diffusion of information: An agent-based modeling approach. *IJATS*, pages 31–46, 2010.

[17] B. Hull et al. CarTel: a distributed mobile sensor computing system. In *SenSys'06*, pages 125–138, 2006.

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

[19] M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, and P. Boda. Peir, the personal environmental impact report, as a platform for participatory sensing systems research. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, MobiSys '09, pages 55–68, New York, NY, USA, 2009. ACM.

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

[21] S. Reddy, D. Estrin, and M. Srivastava. Recruitment framework for participatory sensing data collections. In *Proceedings of the 8th International Conference on Pervasive Computing*, pages 138–155. Springer Berlin Heidelberg, May 2010.

[22] S. Reddy, K. Shilton, G. Denisov, C. Cenizal, D. Estrin, and M. Srivastava. Biketastic: sensing and mapping for better biking. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 1817–1820, New York, NY, USA, 2010. ACM.

[23] 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.

[24] 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.

[25] J. Xie, S. Sreenivasan, G. Korniss, W. Zhang, C. Lim, and B. K. Szymanski. Social consensus through the influence of committed minorities. *CoRR*, abs/1102.3931, 2011.

[26] 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.

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