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

Dong Wang[1], Lance Kaplan[2], Hieu Le[1], Tarek Abdelzaher[1,3]

[1]Department of Computer Science, University of Illinois at Urbana Champaign, Urbana, IL 61801
[2]Networked Sensing and Fusion Branch, US Army Research Labs, Adelphi, MD 20783
[3]Department of Automatic Control, Lund University, Lund, Sweden (Sabbatical Affiliation)

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 reliability 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.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithm

Keywords

Social Sensing, Truth Discovery, Maximum Likelihood Estimation, Expectation Maximization

IPSN'12, April 16–20, 2012, Beijing, China.

1. 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 (e.g., 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 very 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 [19]. 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" [11]. 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.

Prior literature attempted to solve a similar trust analysis problem in information networks using heuristics whose inspiration can be traced back to Google's PageRank [7]. 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 [6, 15, 22, 33, 34]. In contrast, ours is the first attempt to optimally solve the truth discovery problem in social sensing by casting it as one of expectation maximization.

We evaluate our algorithm in simulation, an emulated geotagging scenario as well as a real world social sensing application. 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: we review related work in Section 2. In Section 3, we present the truth discovery model for social sensing applications. The proposed maximum likelihood estimation approach is discussed in Section 4. Implementation and evaluation results are presented in Section 5. We discuss the limitations of current model and future work in Section 6. Finally, we conclude the paper in Section 7.

2. RELATED WORK

Social sensing 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, etc.) and the proliferation of Internet connectivity to upload and share sensed data (e.g., WiFi and 4G networks). A broad overview of social sensing applications is presented in [1]. Some early applications include CenWits [16], a participatory sensor network to rescue hikers in emergency situations, CarTel [18], a vehicular sensor network for traffic monitoring and mitigation, and BikeNet [14], 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 [3, 4]. Social 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 [25], the level of pollution in a city [20], or reporting garbage cans on campus [24]. In addition, 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 [10], information propagation [17] and tipping points [32]. Our paper complements past work by addressing truth discovery in social sensing.

Previous efforts on truth discovery, from the machine learning and data mining communities, provided several interesting heuristics. The Bayesian Interpretation scheme [29] presented an approximation approach to truth estimation that is very sensitive to initial conditions of iterations. Hubs and Authorities [19] 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. extended the fact-finder framework by incorporating prior knowledge into the analysis and proposed several extended algorithms: *Average.Log, Investment, and Pooled Investment* [22]. Yin et al. introduced *TruthFinder* as an unsupervised fact-finder for trust analysis on a providers-facts network [33]. Other fact-finders enhanced the basic framework by incorporating analysis on properties or dependencies within assertions or sources. Galland et al. [15] took the notion of hardness of facts into consideration by proposing their algorithms: *Cosine, 2-Estimates, 3-Estimates*. The source dependency detection problem was discussed and several solutions proposed [6, 12, 13]. Additionally, trust analysis was done both on a homogeneous network [5, 34] and a heterogeneous network [27]. 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. To achieve optimality, 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.

There exists a good amount of literature in machine learning community to improve data quality and identify low quality labelers in a multi-labeler environment. Sheng et al. proposed a repeated labeling scheme to improve label quality by selectively acquiring multiple labels and empirically comparing several models that aggregate responses from multiple labelers [26]. Dekel et al. applied a classification technique to simulate aggregate labels and prune low-quality labelers in a crowd to improve the label quality of the training dataset [9]. However, all of the above approaches made explicit or implicit assumptions that are not appropriate in the social sensing context. For example, the work in [26] assumed labelers were known a priori and could be explicitly asked to label certain data points. The work in [9] assumed most of labelers were reliable and the simple aggregation of their labels would be enough to approximate the ground-truth. In contrast, participants in social sensing usually upload their measurements based on their own obser-

vations and the simple aggregation technique (e.g., majority voting) was shown to be inaccurate when the reliability of participant is not sufficient [22]. The maximum likelihood estimation approach studied in this paper addressed these challenges by intelligently casting the truth discovery problem in social sensing into an optimization problem that can be efficiently solved by the EM scheme.

Our work is related with a type of information filtering system called recommender systems, where the goal is usually to predict a user's rating or preference to an item using the model built from the characteristics of the item and the behavioral pattern of the user [2]. EM has been used in either collaborative recommender systems as a clustering module [21] to mine the usage pattern of users or in a content-based recommender systems as a weighting factor estimator [23] to infer the user context. However, in social sensing, the truth discovery problem targets a different goal: we aim to quantify how reliable a source is and identify whether a measured variable is true or not rather than predict how likely a user would choose one item compared to another. Moreover, users in recommender systems are commonly assumed to provide reasonably good data while the sources in social sensing are in general unreliable and the likelihood of the correctness of their measurements is unknown a priori. There appears no straightforward use of methods in the recommender systems regime for the target problem with unpredictably unreliable data.

3. 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_i represent the i^{th} participant and C_j represent the j^{th} measured variable. S_iC_j denotes 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 the odds of a measured variable to be true given that participant S_i reports it:

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

Let us also define a_i as the (unknown) probability that participant S_i reports a measured variable to be true when it is indeed true, and b_i as the (unknown) probability that participant S_i reports a measured variable to be true when it is in reality false. Formally, a_i and b_i are defined as follows:

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

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

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

The only input to our algorithm is the social sensing topology represented by 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*.

The goal of the algorithm is to compute (i) the best estimate h_j on the correctness of each measured 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:

$$< H^*, E^* > = \underset{<H,E>}{\operatorname{argmax}}\, p(SC | H, E) \tag{4}$$

We also compute the background bias d, which is the overall probability that a randomly chosen measured variable is true. For example, it may represent the probability that any street, in general, is in disrepair. It does not indicate, however, whether any particular claim about disrepair at a particular location is true or not. Hence, one can define the prior of a claim being true as $P(C_j^t) = d$. Note also that, the probability that a participant makes an observation (i.e., s_i) is proportional to the number of measured variables observed by the participant over the total number of measured variables observed by all participants, which can be easily computed from the observation matrix. Hence, one can define the prior $P(S_iC_j) = s_i$. Plugging these, together with t_i into the definition of a_i and b_i, we get the relationship between the terms we defined above:

$$\begin{aligned} a_i &= \frac{t_i \times s_i}{d} \\ b_i &= \frac{(1 - t_i) \times s_i}{1 - d} \end{aligned} \tag{5}$$

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

4. EXPECTATION MAXIMIZATION

In this section, we solve the problem formulated in the previous section using the Expectation-Maximization (EM) algorithm. EM 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 [11]. 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.

4.1 Background

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 θ, then formulate a likelihood function $L(\theta; X, Z) = p(X, Z|\theta)$, such that the maximum likelihood estimate (MLE) of the unknown parameters θ 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 θ for the next iteration.

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

4.2 Mathematical Formulation

Our social 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 C_j such that: $z_j = 1$ when C_j is true and $z_j = 0$ otherwise. We further denote the observation matrix SC as the observed data X, and take $\theta = (a_1, a_2, ...a_M; b_1, b_2, ...b_M; d)$ as the parameter of the model that we want to estimate. The goal is to get the maximum likelihood estimate of θ for the model containing observed data X and latent variables Z.

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

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

where, as we mentioned before, a_i and b_i are the conditional probabilities that participant S_i reports the measured variable C_j to be true given that C_j is true or false (i.e., defined in Equation (2)). $S_i C_j = 1$ when participant S_i reports that C_j is true, and $S_i C_j = 0$ otherwise. d is the background bias that a randomly chosen measured variable is true. Additionally, we assume participants and measured variables are independent respectively. The likelihood function above describes the likelihood to have current observation matrix X and hidden variable Z given the estimation parameter θ we defined.

4.3 Deriving the E-step and M-step

Given the above formulation, substitute the likelihood function defined in Equation (9) into the definition of Q function given by Equation (7) of Expectation Maximization. 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 represents the j^{th} column of the observed SC matrix (i.e., observations of the j^{th} measured variable from all participants) 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 θ, which is given by:

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

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

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

$A(t, j)$ and $B(t, j)$ represent the conditional probability regarding observations about the j^{th} measured variable and current estimation of the parameter θ given the j^{th} measured variable is true or false respectively.

Next we simplify Equation (10) by noting that the conditional probability of $p(z_j = 1|X_j, \theta^{(t)})$ given by Equation (11) is only a function of t and j. Thus, we represent it by $Z(t,j)$. Similarly, $p(z_j = 0|X_j, \theta^{(t)})$ is simply:

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

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

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

The Maximization step (M-step) is given by Equation (8). We choose θ^* (i.e., $(a_1^*, a_2^*, ...a_M^*; b_1^*, b_2^*, ...b_M^*; d^*)$) 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 θ^* that maximizes $Q\left(\theta|\theta^{(t)}\right)$, we set the derivatives $\frac{\partial Q}{\partial a_i} = 0$, $\frac{\partial Q}{\partial b_i} = 0$, $\frac{\partial Q}{\partial d} = 0$ which yields:

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

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

$$\begin{aligned} & \sum_{j \in SJ_i} Z(t,j) \frac{1}{a_i^*} - \sum_{j \in \bar{SJ_i}} Z(t,j) \frac{1}{1 - a_i^*} = 0 \\ & \sum_{j \in SJ_i} (1 - Z(t,j)) \frac{1}{b_i^*} - \sum_{j \in \bar{SJ_i}} (1 - Z(t,j)) \frac{1}{1 - b_i^*} = 0 \\ & \sum_{j=1}^{N} \left[Z(t,j) \frac{1}{d^*} - (1 - Z(t,j)) \frac{1}{1 - d^*}) \right] = 0 \end{aligned} \tag{16}$$

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

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

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. The convergence analysis has been done for EM scheme and it is beyond the scope of this paper [31]. In practice, we can run the algorithm until the difference of estimation parameter between consecutive iterations becomes insignificant. 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^* of participant reliability from the converged values of $a_i^{(t)}$, $b_i^{(t)}$ and $d^{(t)}$ based on their relationship given by Equation (5). This completes the mathematical development. We summarize the resulting algorithm in the subsection below.

4.4 The Final Algorithm

In summary of the EM scheme derived above, the input is the observation matrix SC from social sensing data, and the output is the maximum likelihood estimation of participant reliability and measured variable correctness (i.e., E^* and H^* vector defined in Equation (4)). In particular, given the observation matrix SC, our algorithm begins by initializing the parameter θ[2]. The algorithm then performs the E-steps and M-steps iteratively until θ converges. Specifically, we compute the conditional probability of a measured variable to be true (i.e., $Z(t,j)$) from Equation (11) and the estimation parameter (i.e., $\theta^{(t+1)}$) from Equation (17). After the estimated value of θ 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 $\theta^{(t)}$ (i.e., a_i^c, b_i^c and d^c) based on Equation (5) as shown in the pseudocode of Algorithm 1.

One should note that a theoretical quantification of accuracy of maximum likelihood estimation (MLE) using the EM scheme is well-known in literature, and can be done using the Cramer-Rao lower bound (CRLB) on estimator variance [8]. In estimation theory, if the estimation variance of an unbiased estimator reaches the Cramer-Rao lower bound, the estimator provides the maximum likelihood estimation and the CRLB quantifies the minimum estimation variance. The estimator proposed in this paper is shown to operate

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

Algorithm 1 Expectation Maximization Algorithm

Initialize θ ($a_i = s_i, bi = 0.5 \times s_i$, d =Random number in $(0, 1)$)
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 $a_i^{(t+1)}, b_i^{(t+1)}, d^{(t+1)}$ based on Equation (17)
    update $a_i^{(t)}, b_i^{(t)}, d^{(t)}$ with $a_i^{(t+1)}, b_i^{(t+1)}, d^{(t+1)}$ 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 $b_i^{(t)}$; d^c = converged value of $d^{(t)}$
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, b_i^c and d^c based on Equation (5)
end for
Return the computed optimal estimates of measured variables $C_j = h_j^*$ and source reliability e_i^*.

at this bound and hence reach the maximum likelihood estimation [30]. This observation makes it possible to quantify estimation accuracy, or confidence in results generated from our scheme, using the Cramer-Rao lower bound.

5. EVALUATION

In this section, we carry out 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 both an emulated participatory sensing scenario and a real world social sensing application. We show that the new algorithm outperforms the state of the art.

5.1 A Simulation Study

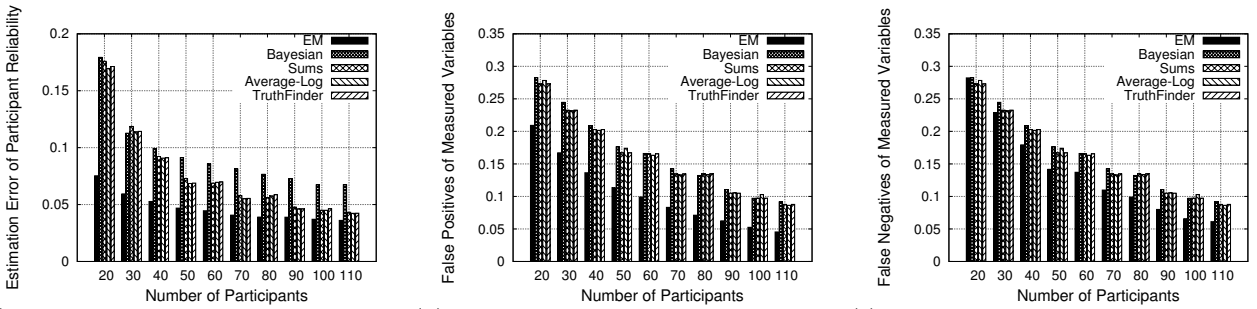
We built a simulator in Matlab 7.10.0 that generates a random number of participants and measured variables. A random probability P_i is assigned to each participant S_i representing his/her reliability (i.e., the ground truth probability that they report correct observations). For each participant S_i, L_i observations are generated. Each observation has a probability t_i of being true (i.e., reporting a variable as true correctly) and a probability $1-t_i$ 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 t_i be uniformly distributed between 0.5 and 1 in our experiments[3]. For initialization, the initial values of participant reliability (i.e., t_i) in the evaluated schemes are set to the mean value of its definition range.

In recent work, a heuristic called *Bayesian Interpretation* was demonstrated to outperform all contenders from prior literature [29]. Bayesian Interpretation takes a linear approximation approach to convert the credibility ranks of fact-finders into a Bayesian probability that a participant reports correctly or the measured variable is true. In Bayesian Interpretation, the performance evaluation results were averaged over multiple observation matrices for a given participant reliability distribution. This is intended to approximate performance where highly connected sensing topologies are available (e.g., observations from successive time intervals involving the same set of sources and measured variables). In this paper, we consider more challenging conditions not investigated in [29], where only a *single observation matrix* is taken as the input into the algorithm. This is intended to understand the algorithm's performance in more realistic scenarios where the sensing topologies are sparsely connected. We compare EM to Bayesian Interpretation and three state-of-art fact-finder schemes from prior literature that can function using only the inputs offered in our problem formulation [19, 22, 33]. Results show a significant performance improvement of EM over all heuristics compared.

In the first experiment, we compare the estimation accuracy of EM and the baseline schemes 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. To favor our competition, we "cheat" by giving the other algorithms the correct value of bias d (in this case, $d = 0.5$). The average number of observations per participant was set to 100. The number of participants was varied from 20 to 110. Reported results are averaged over 100 random participant reliability distributions. Results are shown in Figure 1. 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.

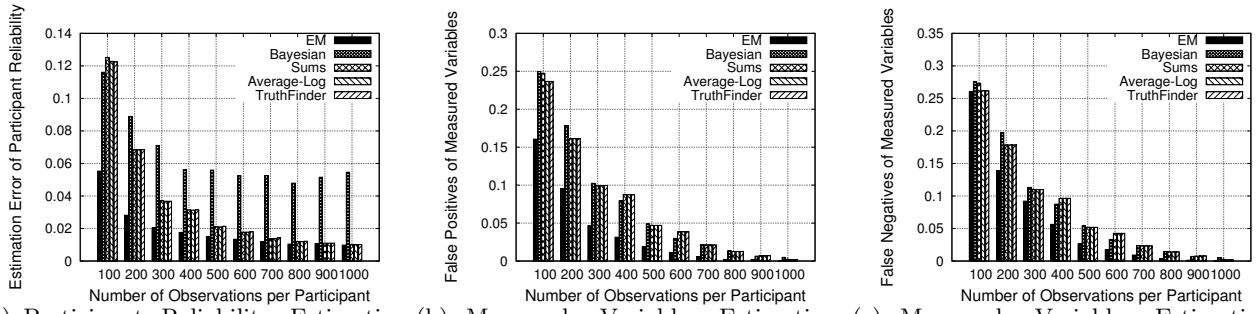
The second experiment compares EM with baseline schemes when the average number of observations per participant changes. As before, we fix the number of correctly and incorrectly reported variables to 1000 respectively. Again, we favor our competition by giving their algorithms the correct value of background bias d (here, $d = 0.5$). 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. The results are shown in Figure 2. 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 number of observations per participant is small and starts to gain advantage as the average number of observations per par-

[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



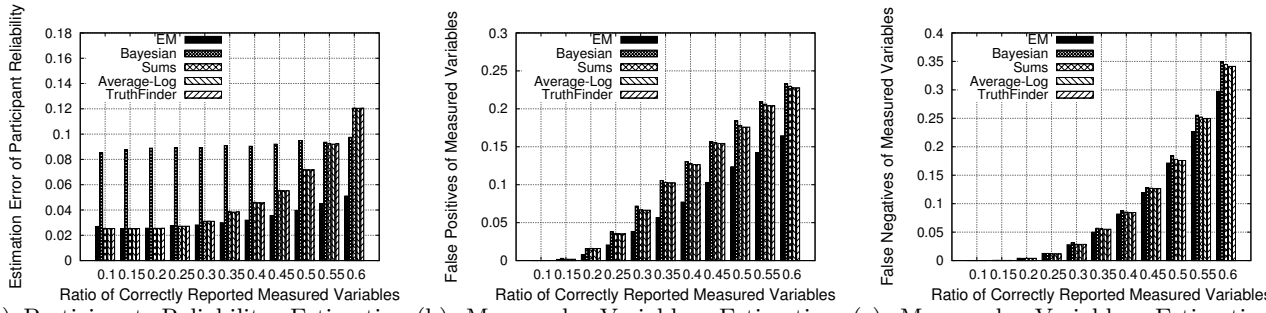
(a) Participant Reliability Estimation Accuracy (b) Measured Variable Estimation: False Positives (c) Measured Variable Estimation: False Negatives

Figure 1: Estimation Accuracy versus Number of Participants



(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



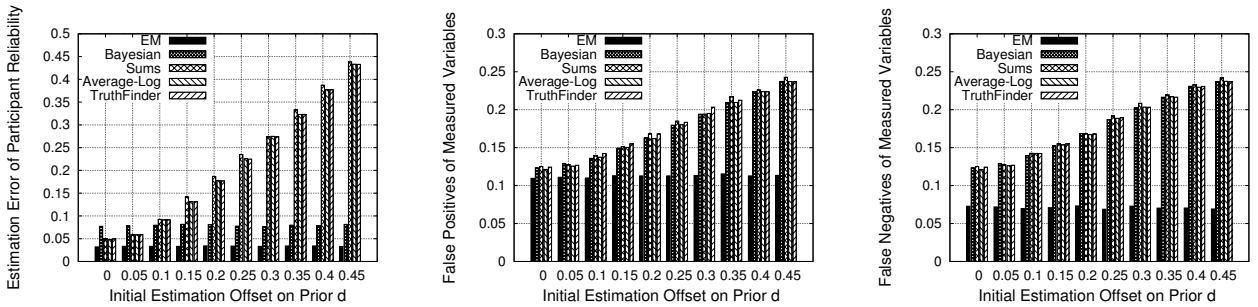
(a) Participant Reliability Estimation Accuracy (b) Measured Variable Estimation: False Positives (c) Measured Variable Estimation: False Negatives

Figure 3: Estimation Accuracy versus Ratio of Correctly Reported Measured Variables

ticipant 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.

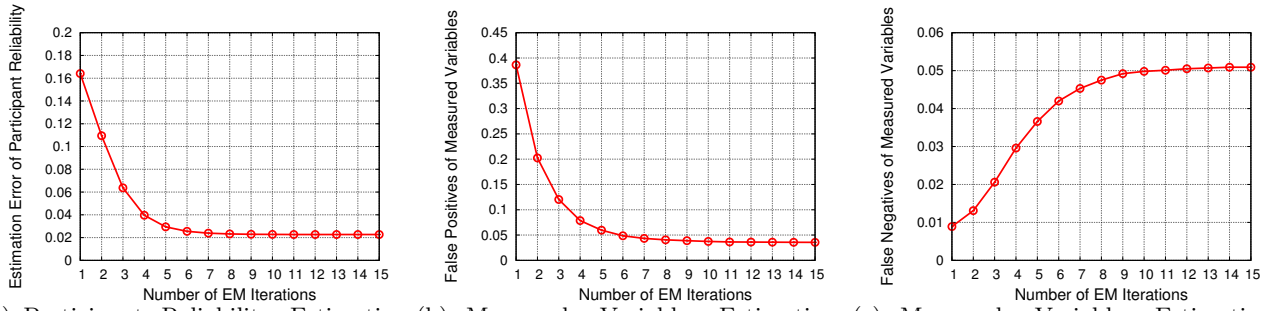
The third experiment examines the effect of changing the measured variable mix on the estimation accuracy of all schemes. We vary the ratio of the number of correctly reported variables to the total number of reported variables from 0.1 to 0.6, while fixing the total number of such variables to 2000. To favor the competition, the background bias d is given correctly to the other algorithms (i.e., $d = varying\ \ ratio$). 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. We observe that EM has almost the same performance as other fact-finder baselines when the fraction of correctly reported variables is relatively small. The reason is that the small amount of true measured variables are densely observed and most of them can be easily differentiated from the false ones by both EM and baseline fact-finders. 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. Throughout the first to the third experiments, we also observe that the Bayesian interpretation scheme predicts less accurately than other heuristics. This is because the estimated posterior probability of a participant to be reliable or a measured variable to be true in Bayesian interpretation is a linear transform of the participant's or the measured variable's credibility values. Those values obtained from a single or sparse observation matrix may not be very accurate and refined [29].

The fourth experiment evaluates the performance of EM and other schemes when the offset of the initial estimation on the background bias d varies. The offset is defined as the difference between initial estimation on d and its ground-



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

Figure 4: Estimation Accuracy versus Initial Estimation Offset on Prior d



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

Figure 5: Convergence Property of the EM Algorithm

truth. We fix the number of correctly and incorrectly reported variables to 1000 respectively (i.e., $d = 0.5$). We vary the absolute value of the initial estimate offset on d from 0 to 0.45. The reported results are averaged for both positive and negative offsets of the same absolute value. The number of participants is fixed at 50 and the average number of observations per participant is set to 150. Reported results are averaged over 100 experiments. Figure 4 shows the results. We observe that the performance of EM scheme is stable as the offset of initial estimate on d increases. On the contrary, the performance of other baselines degrades significantly when the initial estimate offset on d becomes large. This is because the EM scheme incorporates the d as part of its estimation parameter and provides the MLE on it. However, other baselines depend largely on the correct initial estimation on d (e.g., from the past history) to find out the right number of correctly reported measured variables. These results verify the robustness of the EM scheme when the accurate estimate on the prior d is not available to obtain.

The fifth experiment shows the convergence property of the EM iterative algorithm in terms of the estimation error on participant reliability, as well as the false positives and false negatives on measured variables. We fix the number of correctly and incorrectly reported variables to 1000 respectively and set the initial estimate offset on d to 0.3. The number of participants is fixed at 50 and the average number of observations per participant is set to 250. Reported results are averaged over 100 experiments. Figure 5 shows the results. We observe that both the estimation error on participant reliability and false positives/negatives on measured variable converge reasonably fast (e.g., less than 10 iterations) to stable values as the number of iterations of EM algorithm increases. It verifies the efficiency of applying EM scheme to solve the maximum likelihood estimation problem formulated.

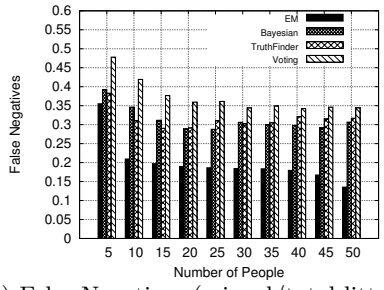
This concludes our general simulations. In the next subsection, we emulate the performance of a specific social sensing application.

5.2 A Geotagging Case Study

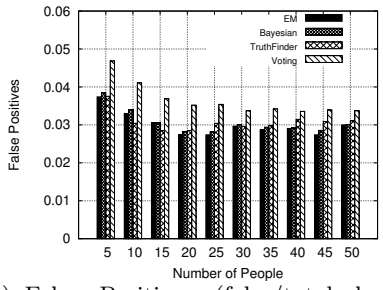
In this subsection, we applied the proposed EM scheme to a typical social sensing application: Geotagging locations of litter in a park or hiking area. In this application, litter may be found along the trails (usually proportionally to their popularity). Participants visiting the park geotag and report 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 as the ratio of litter locations missed by a 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 a scheme, 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 as shown in Figure 6. 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

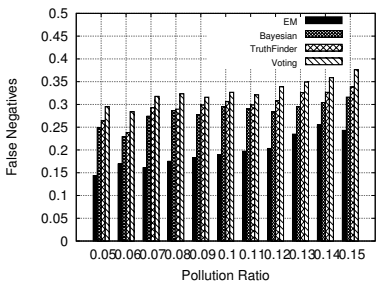


(a) False Negatives (missed/total litter)

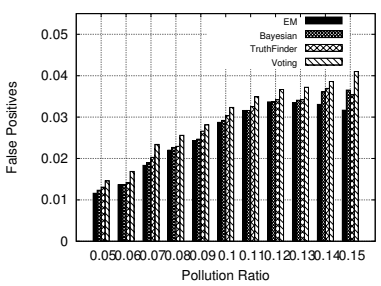


(b) False Positives (false/total locations)

Figure 7: Litter Geotagging Accuracy versus Number of People Visiting the Park

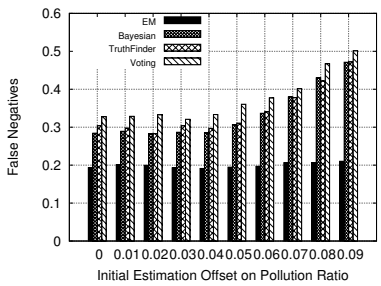


(a) False Negatives (missed/total litter)

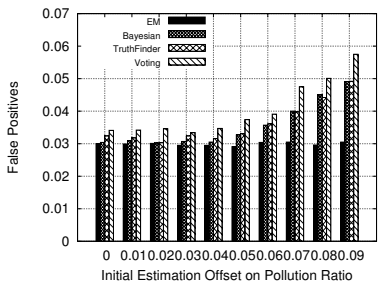


(b) False Positives (false/total locations)

Figure 8: Litter Geotagging Accuracy versus Pollution Ratio of the Park



(a) False Negatives (missed/total litter)



(b) False Positives (false/total locations)

Figure 9: Litter Geotagging Accuracy versus Initial Estimation Offset on Pollution Ratio of Park

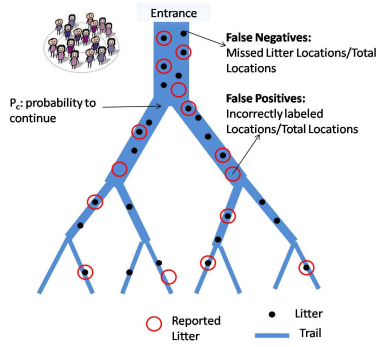


Figure 6: A Simplified Trail Map of Geotagging Application

discretely labeled locations (e.g., numbered distance markers). In our emulation, 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.

In the third experiment, we evaluate the effect of the initial estimation offset of the pollution ratio on the performance of different schemes. The pollution ratio is fixed at 0.1 and the number of individuals visiting the park is set to be 40. We vary the absolute value of initial estimation offset of the pollution ratio from 0 to 0.09. Results are averaged over both positive and negative offsets of the same absolute value. The estimation results of different schemes are shown in Figure 9. Observe that EM finds more actual litter locations and reports less falsely labeled locations than other baselines as the initial estimation offset of pollution ratio increases. Additionally, the performance of EM scheme is stable while the performance of other baselines drops substantially when the initial estimation offset of the pollution ratio becomes large.

The above evaluation demonstrates that the new EM scheme generally outperforms the current state of the art in inferring facts from social sensing data. This is because the state of the art heuristics infer the reliability of participants and correctness of facts based on the hypothesis that their relationship can be approximated *linearly* [22, 29, 33]. However, EM scheme makes its inference based on a maximum likelihood hypothesis that is most consistent with the observed sensing data, thus it provides an optimal solution.

5.3 A Real World Application

In this subsection, we evaluate the performance of the proposed EM scheme through a real-world social sensing application, based on Twitter. The objective was to see whether our scheme would distill from Twitter feeds important events that may be newsworthy and reported by media. Specifically, we followed the news coverage of Hurricane Irene and manually selected, as ground truth, 10 important events reported by media during that time. Independently from that collection, we also obtained more than 600,000 tweets originating from New York City during Hurricane Irene using the Twitter API (by specifying keywords as "hurricane", "Irene" and "flood", and the location to be New York). These tweets were collected from August 26 until September 2nd, roughly when Irene struck the east coast. Retweets were removed from the collected data to keep sources as independent as possible.

We then generated an observation matrix from these tweets by clustering them based on the Jaccard distance metric (a simple but commonly used distance metric for micro-blog data [28]). Each cluster was taken as a statement of claim about current conditions, hence representing a measured variable in our model. Sources contributing to the cluster were connected to that variable forming the observation matrix. In the formed observation matrix, participants are the twitter users who provided tweets during the observation period, measured variables are represented by the clusters of tweets and the element S_iC_j is set to 1 if the tweets of participant S_i belong to cluster C_j, or to 0 otherwise. The matrix was then fed to our EM scheme. We ran the scheme on the collected data and picked the top (i.e., most credible) tweet in each hour. We then checked if our 10 "ground truth" events were reported among the top tweets. Table 1 compares the ground truth events to the corresponding top hourly tweets discovered by EM. The results show that indeed all events were reported correctly, demonstrating the value of our scheme in distilling key important information from large volumes of noisy data.

#	Media	Tweet found by EM
1	East Coast Braces For Hurricane Irene; Hurricane Irene is expected to follow a path up the East Coast	@JoshOchs A #hurricane here on the east coast
2	Hurricane Irene's effects begin being felt in NC, The storm, now a Category 2, still has the East Coast on edge.	Winds, rain pound North Carolina as Hurricane Irene closes in http://t.co/0gVOSZk
3	Hurricane Irene charged up the U.S. East Coast on Saturday toward New York, shutting down the city, and millions of Americans sought shelter from the huge storm.	Hurricane Irene rages up U.S. east coast http://t.co/u0XiXow
4	The Wall Street Journal has created a way for New Yorkers to interact with the location-based social media app Foursquare to find the nearest NYC hurricane evacuation center.	Mashable - Hurricane Irene: Find an NYC Evacuation Center on Foursquare ... http://t.co/XMtpH99
5	Following slamming into the East Coast and knocking out electricity to more than a million people, Hurricane Irene is now taking purpose on largest metropolitan areas in the Northeast.	2M lose power as Hurricane Irene moves north - Two million homes and businesses were without power ... http://t.co/fZWkEU3
6	Irene remains a Category 1, the lowest level of hurricane classification, as it churns toward New York over the next several hours, the U.S. National Hurricane Center said on Sunday.	Now its a level 1 hurricane. Let's hope it hits NY at Level 1
7	Blackouts reported, storm warnings issued as Irene nears Quebec, Atlantic Canada.	DTN Canada: Irene forecast to hit Atlantic Canada http://t.co/MjhmeJn
8	President Barack Obama declared New York a disaster area Wednesday, The New York Times reports, allowing the release of federal aid to the state's government and individuals.	Hurricane Irene: New York State Declared A Disaster Area By President Obama
9	Hurricane Irene's rampage up the East Coast has become the tenth billion-dollar weather event this year, breaking a record stretching back to 1980, climate experts said Wednesday.	Irene is 10th billion-dollar weather event of 2011.
10	WASHINGTON- On Sunday, September 4, the President will travel to Paterson, New Jersey, to view damage from Hurricane Irene.	White House: Obama to visit Paterson, NJ Sunday to view damage from Hurricane Irene

Table 1: Ground truth events and related tweets found by EM in Hurricane Irene

6. DISCUSSION AND FUTURE WORK

Participants (sources) are assumed to be independent from each other in the current EM scheme. However, sources can sometimes be dependent. That is, they copy observations from each other in real life (e.g., retweets of Twitter). Regarding possible solutions to this problem, one possibility is to remove duplicate observations from dependent sources and only keep the original ones. This can be achieved by applying copy detection schemes between sources [12, 13]. Another possible solution is to cluster dependent sources based on some *source-dependency* metric [6]. In other words, sources in the same cluster are closely related with each other but independent from sources in other clusters. Then we can apply the developed algorithm on top of the clustered sources.

Observations from different participants on a given measured variable are assumed to be *corroborating* in this paper.

This happens in social sensing applications where people do not report "lack of problems". For example, a group of participants involved in a geotagging application to find litter of a park will only report locations where they observe litter and ignore the locations they don't find litter. However, sources can also make conflicting observations in other types of applications. For example, comments from different reviewers in an on-line review system on the same product often contradict with each other. Fortunately, our current model can be flexibly extended to handle conflicting observations. The idea is to extend the estimation vector to incorporate the conflicting states of a measured variable and rebuild the likelihood function based on the extended estimation vector. The general outline of the EM derivation still holds.

The current EM scheme is mainly designed to run on static data sets, where the computation overhead stays reasonable even when the dataset scales up (e.g., the Irene dataset). However, such computation may become less efficient for streaming data because we need to re-run the algorithm on the whole dataset from scratch every time the dataset gets updated. Instead, it will be more technically sound that the algorithm only runs on the updated dataset and combines the results with previously computed ones in an optimal (or suboptimal) way. One possibility is to develop a scheme that can compute the estimated parameters of interest recursively over time using incoming measurements and a mathematical process model. The challenge here is that the relationship between the estimation from the updated dataset and the complete dataset may not be linear. Hence, linear regression might not be generally plausible. Rather, recursive estimation schemes, such as the Recursive Bayesian estimation, would be a better fit. The authors are currently working on accommodating the above extensions.

7. 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.

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.

8. REFERENCES

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

[2] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 17(6):734–749, 2005.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] H. Cramer. *Mathematical Methods of Statistics*. Princeton Univ. Press., 1946.

[9] O. Dekel and O. Shamir. Vox populi: Collecting high-quality labels from a crowd. In *In Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

[10] 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.

[11] 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.

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

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

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

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

[16] 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.

[17] 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.

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

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

[20] 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.

[21] N. Mustapha, M. Jalali, and M. Jalali. Expectation maximization clustering algorithm for user modeling in web usage mining systems. *European Journal of Scientific Research*, 32(4):467–476, 2009.

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

[23] D. Pomerantz and G. Dudek. Context dependent movie recommendations using a hierarchical bayesian model. In *Proceedings of the 22nd Canadian Conference on Artificial Intelligence: Advances in Artificial Intelligence*, Canadian AI '09, pages 98–109, Berlin, Heidelberg, 2009. Springer-Verlag.

[24] 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.

[25] 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.

[26] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 614–622, New York, NY, USA, 2008. ACM.

[27] 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.

[28] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. 2005.

[29] 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.

[30] 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.

[31] C. F. J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[32] 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.

[33] 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.

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