Maximum Likelihood Analysis of Conflicting Observations in Social Sensing

DONG WANG, University of Illinois at Urbana-Champaign
LANCE KAPLAN, US Army Research Laboratory
TAREK F. ABDELZAHER, University of Illinois at Urbana-Champaign

This article 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. 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, we pose 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. In our previous conference paper, we offered the first maximum likelihood solution to the above truth discovery problem for *corroborating observations* only. In contrast, this paper extends the conference paper and provides the first maximum likelihood solution to handle the cases where measurements from different participants may be *conflicting*. The paper focuses on binary measurements. The approach is shown to outperform our previous work used for corroborating observations, the state of the art fact-finding baselines, as well as simple heuristics such as majority voting.

Categories and Subject Descriptors: H.4 [**Information Systems Applications**]: Miscellaneous

General Terms: Algorithm

Additional Key Words and Phrases: Social Sensing, Truth Discovery, Conflicting Observations, Maximum Likelihood Estimation, Expectation Maximization

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 inter-

This journal paper is a significantly enhanced version of a preliminary work that was presented at ACM/IEEE IPSN'12. Research reported in this paper was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.
Author's addresses: D. Wang, T. F. Abdelzaher, Computer Science Department, University of Illinois at Urbana-Champaign. L. Kaplan, Networked Sensing & Fusion Branch, US Army Research Laboratory.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© YYYY ACM 1550-4859/YYYY/01-ARTA $15.00
DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

est. 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 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) mutually exclusive measurements (e.g., reporting whether or not litter exists at each of multiple locations of interest). The observations from participants can be either corroborating or conflicting. In our previous conference paper, we developed a maximum likelihood estimator to optimally solve the above problem where only corroborating observations exist (e.g., people only report existence of problem and ignore lack of problems) [Wang et al. 2012a]. In contrast, this paper extended the basic model and maximum likelihood estimation (MLE) approach used in the conference paper to solve the truth discovery problem in more challenging cases where observations from participants may be *conflicting*. The new algorithm makes inferences regarding both source reliability and measurement correctness by observing which observations *coincide* and which *contradict* and assigns truth values to measurements without prior knowledge of source reliability. Our approach is shown to be very accurate in estimating participant reliability and assessing measurement correctness in the context of conflicting observations from different participants.

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 [Kleinberg 1999]. 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" [Dempster et al. 1977]. 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 similar trust analysis problem in information network using heuristics whose inspiration can be traced back to Google's PageRank [Brin and Page 1998]. 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 [Pasternack and Roth 2010; Yin et al. 2008; Galland et al. 2010; Berti-Equille et al. 2009; Yin and Tan 2011; Zhao et al. 2012]. In contrast, our prior conference pub-

lication [Wang et al. 2012a] is the first attempt to optimally solve the truth discovery problem in social sensing where the observations from different participants are corroborating by casting it as an expectation maximization problem. This paper extends the model and maximum likelihood estimation approach used in the conference paper and proposed an optimal solution to the truth discovery problem where the observations may be *conflicting*.

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 scheme for conflicting observations in this paper outperforms our previous work [Wang et al. 2012a] used for corroborating observations, the state-of-art fact-finding 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 propose the truth discovery model for social sensing applications with conflicting observations and the new maximum likelihood estimation (MLE) approach (the EM-Conflict scheme) as the solution. Section 4 presents truth discovery model and MLE approach for corroborating observations as a special case of the one discussed in Section 3. Implementation and evaluation results are presented in Section 5. We discuss the limitations of current model 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 common 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 [Abdelzaher et al. 2007]. Some early applications include CenWits [Huang et al. 2005], CarTel [Hull et al. 2006] and BikeNet [Eisenman et al. 2007]. More recent work has focused on addressing new challenges emerging in social sensing applications such as preserving privacy of participants [Ahmadi et al. 2010; Pham et al. 2010], improving energy efficiency of sensing devices [Nath 2012; Park et al. 2011] and building general models in sparse and multi-dimensional social sensing space [Ahmadi et al. 2011; Wang et al. 2011b]. Examples include privacy-aware regression modeling, a data fusion technique that produce the same model as that computed from raw data by properly computing non-invertible aggregates of samples [Ahmadi et al. 2010]. Authors in [Pham et al. 2010] gave special attention to preserving privacy over time series data based on the observation that sensor data stream typically comprises a correlated series of sampled data from some continuous physical phenomena. Acquisitional Context Engine (ACE) is a middleware that infers the unknown human activity attribute from known ones by exploiting the observation that the values of various human context attribute are limited by physical constraints and hence highly correlated [Nath 2012]. E-Gesture is an energy efficient gesture recognition architecture that significantly reduces the energy consumption of mobile sensing device while keeping the recognition accuracy acceptable [Park et al. 2011]. Sparse regression cube is a modeling technique that combines estimation theory and data mining techniques to enable reliable modeling at multiple degrees of abstraction of sparse social sensing data [Ahmadi et al. 2011]. A further improved model to consider the data collection cost was proposed in [Wang et al. 2011b]. Moreover, 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 for some mutual interests. Examples include documenting the quality of roads [Reddy et al. 2010b], reporting garbage cans on campus [Reddy et al. 2010a] or predicting the

bus arrival time at various bus stops [Zhou et al. 2012]. In addition, social sensing can also be triggered spontaneously without prior coordination. Examples include modeling human mobility patterns in different metropolitan areas [Becker et al. 2012], predicting the expected fare and duration of the taxi ride in large cities [Balan et al. 2011] or real-time summarizing scheduled events from twitter streams [Zubiaga et al. 2012]. 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 [Delre et al. 2007], information propagation [Hui et al. 2010] and tipping points [Xie et al. 2011]. 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. Hubs and Authorities [Kleinberg 1999] used a basic fact-finder where the belief in a claim 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 asserting a given claim and the claims asserted 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* [Pasternack and Roth 2010]. Yin et al. introduced *TruthFinder* as an unsupervised fact-finder for trust analysis on a providers-facts network [Yin et al. 2008]. Other fact-finders enhanced the basic framework by incorporating analysis on properties or dependencies within claims or sources. Galland et al. [Galland et al. 2010] 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 [Berti-Equille et al. 2009; Dong et al. 2009; Dong et al. 2010]. More recent works have adapted the Bayesian analysis to model the source trustworthiness in an explicit and probabilistic way and improved the accuracy of truth estimation. Wang et al. [Wang et al. 2011a] proposed the Bayesian Interpretation scheme as an approximation approach to correctly quantify the conclusions obtained from the basic fact-finding scheme. Zhao et al. [Zhao et al. 2012] presented another Bayesian based approach to model different types of errors made by sources and merge multi-valued attribute types of entities in data integration systems. Additionally, trust analysis was done both on a homogeneous network [Balakrishnan 2011; Yin and Tan 2011] and a heterogeneous network [Sun et al. 2009]. The EM scheme proposed in our recent work [Wang et al. 2012a] was the first that finds a maximum likelihood estimator to directly and optimally quantify the accuracy of conclusions obtained from credibility analysis in social sensing where observations from participants are *corroborating*. In contrast, this paper extended the previous EM model to handle a more challenging case where observations from participants may be *conflicting*. To achieve optimality, we intentionally start with a simplified application model, where 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 vast literature in the 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 [Sheng et al. 2008]. 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 [Dekel and Shamir 2009]. 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 [Sheng et al. 2008] assumed

labelers were known a priori and could be explicitly asked to label certain data points. The work in [Dekel and Shamir 2009] 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 observations and the simple aggregation technique (e.g., majority voting) was shown to be inaccurate when the reliability of participant is not sufficient [Pasternack and Roth 2010]. The maximum likelihood estimation approach studied in this paper addresses 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 system, 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 [Adomavicius and Tuzhilin 2005]. EM has been used in either collaborative recommender system as a clustering module [Mustapha et al. 2009] to mine the usage pattern of users or in a content-based recommender system as a weighting factor estimator [Pomerantz and Dudek 2009] 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 claim is true or not rather than predict how likely a user would choose one item compared to another. Moreover, users in recommender system 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 recommendation system regime for the target problem with unpredictably unreliable data.

Several previous efforts on data cleaning and outlier analysis from data mining and noise removal from statistics addressed some notion of noisy data [Duda et al. 2001; Inc and Staff 1997; Johnson and Wichern 2002; J.Han et al. 2011; Kalman 1960; Doucet et al. 2001]. They differ in the assumptions made, the modeling approaches applied and the objectives targeted at. For example, Bayesian inference and decision tree induction techniques are applied to fill the missing values of data by predictions from their constructed model [Duda et al. 2001]. Binning and linear regression techniques are used to smooth the noisy data by either using bin means or fitting data into some linear functions [Inc and Staff 1997; Johnson and Wichern 2002]. Clustering techniques are widely used to detect outliers by organizing similar data values into clusters and identifying the ones that fall outside the clusters as outliers [J.Han et al. 2011]. Other approaches are used in statistics to filter noises from continuous data [Kalman 1960; Doucet et al. 2001]. Kalman filter is an efficient reclusive filter that estimates some latent variables of a linear dynamic system from a series of noisy measurements [Kalman 1960]. It produces estimates of the measurements by computing a weighted average of the predicted values based on their uncertainty. Particle filters are more sophisticated filters that are based on Sequential Monte Carlo methods. They are often used to determine the distribution of a latent variable whose state space is not restricted to Gaussian distribution [Doucet et al. 2001]. Our work is complementary to the above efforts. On one hand, an appropriately cleaned and outlier-removed dataset will likely result in a better estimation of our scheme. On the other hand, outliers or noises may not be completely (or even possibly) removed by the data cleaning and outlier analysis techniques mentioned above due to their own limitations (e.g., linear model assumption, continuous data assumption, known data distribution assumption and etc.). The quantifiable estimation on both information source and observed data provided by our approach could actually help the data cleaning and outlier analysis tools do a better job.

3. TRUTH DISCOVERY MODEL FOR SOCIAL SENSING WITH CONFLICTING OBSERVATIONS

3.1. Truth Discovery Model for Conflicting Observations

To formulate the truth discovery problem in social sensing in a manner amenable to rigorous optimization, we consider a basic social sensing application model where a group of M participants (sources), $S_1, ..., S_M$, make individual observations about a set of N claims $C_1, ..., C_N$ in their environment. For example, a group of individuals interested in the appearance of a park in their neighborhood might join a sensing campaign to report the litter locations of the park. Hence, each claim denotes the existence or lack thereof of a litter at a given location[1]. The reported observations from different participants on the same claim may be *conflicting* (e.g., some people report litter exists at location X while others report X to be clean). In general, the claim is assumed to have K mutually exclusive possible values and only one of them represents the true value of the claim. In the model to handle conflicting observations, we assume that observations from participants assert one of the K values of the corresponding claim and can be conflicting with each other. Let S_i represent the i^{th} participant and C_j represent the j^{th} claim. Each participant generally observes only a subset of all claims (e.g., the conditions at locations they visited). Let $S_iC_j = k$ denote participant S_i reports the claim C_j to be of value k for $k = 0, ..., K$. Note that $S_iC_j = 0$ means that participant S_i does not report an observation for claim C_j. Let the probability that participant S_i reports the claim to be of value k be s_i^k (i.e, $s_i^k = P(S_iC_j = k)$ for $k = 1, ..., K$). Let $s_i^{\bar{k}}$ represent the probability that S_i reports a claim to be of value other than k (i.e., $s_i^{\bar{k}} = \sum_{k' \neq 0,k} s_i^{k'}$).

Further, t_i denotes the probability that participant S_i is right (i.e., probability that the participant's observation *matches* the ground truth of the claim) and $1 - t_i$ denotes the probability that it is wrong. Note that, this probability depends on the participant's reliability, which is not known *a priori*. Our goal is to determine the true value of each claim as well as the reliability of each participant. 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 us also define $a_{k,i}^T$ and $a_{k,i}^F$ as the (unknown) probability that participant S_i reports a claim to be of value k and value other than k when the claim is indeed of value k respectively. Formally, $a_{k,i}^T$ and $a_{k,i}^F$ are defined as follows:

$$\begin{aligned} a_{k,i}^T &= P(S_iC_j = k | C_j = k) \\ a_{k,i}^F &= \sum_{k' \neq 0,k}^{K} P(S_iC_j = k' | C_j = k) \end{aligned} \tag{1}$$

where $C_j = k$ denotes the claim C_j is indeed of value k for $k = 1, ..., K$. We assume that participant S_i can report one (and only one) of the K mutually exclusive values for claim C_j (i.e., a source is not self-contradictory on its assertion for a claim). Since a source may not assert a claim ($k = 0$), $a_{k,i}^T + a_{k,i}^F \leq 1$.

Let us define the observation matrix SC to handle conflicting observations: $S_iC_j = k$ when participant S_i reports that C_j is of value k, $S_iC_j = 0$ when no reports about C_j from S_i. Let us call this observation matrix the *conflicting observation matrix*. Let d_k represent the overall prior probability that an arbitrary claim is of value k.

[1]We assume that locations are discretized, and therefore finite (e.g., they are given by mile markers.)

Plugging these, together with t_i into the definition of $a_{k,i}^T$ and $a_{k,i}^F$, we get the relations between the terms defined above by using the Bayesian theorem:

$$a_{k,i}^T = \frac{t_i \times s_i^k}{d_k}$$
$$a_{k,i}^F = \frac{(1-t_i) \times s_i^{\bar{k}}}{d_k} \tag{2}$$

3.2. Expectation Maximization for Conflicting Observations

In this subsection, we solve the problem formulated in the previous subsection 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 [Dempster et al. 1977]. Intuitively, EM iteratively "completes" the data by "guessing" the values of hidden variables and then re-estimates the parameters by using the guessed values as true values.

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:

$$L(\theta; X) = p(X|\theta) = \sum_Z p(X, Z|\theta) \tag{3}$$

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.

$$Q\left(\theta|\theta^{(n)}\right) = E_{Z|X,\theta^{(n)}}[\log L(\theta; X, Z)] \tag{4}$$

— 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.

$$\theta^{(n+1)} = \underset{\theta}{\operatorname{argmax}}\, Q\left(\theta|\theta^{(n)}\right) \tag{5}$$

Our social sensing problem fits nicely into the Expectation Maximization (EM) model. First, we introduce a latent variable Z for each claim to indicate the value of the claim. Specifically, we have a corresponding variable z_j for the j^{th} claim C_j such that: $z_j = k$ when C_j is of value k. We further denote the observation matrix SC as the observed data X, and take $\theta = (\theta_1, \theta_2, ..., \theta_K)$ where $\theta_k = (a_{k,1}^T, a_{k,1}^F, a_{k,2}^T, a_{k,2}^F ... a_{k,M}^T, a_{k,M}^F, d_k)$ as the parameters 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.

Given the estimation parameter and hidden variables defined above, the likelihood function $L(\theta; X, Z)$ for conflicting observations is:

$$\begin{aligned} L(\theta; X, Z) &= p(X, Z|\theta) \\ &= \prod_{j=1}^{N} \Bigg\{ \sum_{k=1}^{K} \Big[\prod_{i=1}^{M} {a_{k,i}^{T}}^{S_i C_j^k} \times {a_{k,i}^{F}}^{S_i C_j^{\bar{k}}} \\ &\times (1 - a_{k,i}^{T} - a_{k,i}^{F})^{(1 - S_i C_j{}^k - S_i C_j{}^{\bar{k}})} \times d_k \times z_j^k \Big] \Bigg\} \end{aligned} \tag{6}$$

where $S_i C_j^k = 1$ when participant S_i asserts the claim C_j to be of value k (i.e., $S_i C_j = k$) and 0 otherwise, $S_i C_j^{\bar{k}} = 1$ when participant S_i asserts the claim C_j to be of value other than k (i.e., $S_i C_j \neq k$ or 0) and 0 otherwise, and $z_j^1, z_j^2, ..., z_j^K$ is a set of indicator variables for claim C_j where $z_j^k = 1$ when C_j is of value k and $z_j^k = 0$ otherwise. Additionally, the values of $S_i C_j$ are statistically independent over the M participants and N claims. The likelihood function above describes the likelihood to have current observation matrix X and hidden variable Z given the estimation parameter θ we defined.

Given the above formulation, we can derive the E-Step as

$$\begin{aligned} &Q\left(\theta|\theta^{(n)}\right) = \\ &\sum_{j=1}^{N} \Bigg\{ \sum_{k=1}^{K} Z_k(n,j) \times \Big[\sum_{i=1}^{M} \Big(S_i C_j^k \log a_{k,i}^{T} + S_i C_j^{\bar{k}} \log a_{k,i}^{F} \\ &+ (1 - S_i C_j^k - S_i C_j^{\bar{k}}) \log(1 - a_{k,i}^{T} - a_{k,i}^{F}) + \log d_k \Big) \Big] \Bigg\} \end{aligned} \tag{7}$$

where $Z_k(n, j)$ is given by:

$$\begin{aligned} Z_k(n,j) =& p(z_j = k|X_j, \theta^{(n)}) \\ &= \frac{A_k(n,j) \times d_k^{(n)}}{\sum_{k=1}^{K} A_k(n,j) \times d_k^{(n)}} \end{aligned} \tag{8}$$

where $A_k(n, j)$ is defined as:

$$\begin{aligned} A_k(n,j) &= p(X_j, \theta^{(n)}|z_j = k) \\ &= \prod_{i=1}^{M} \Big\{ {a_{k,i}^{T}}^{(n) S_i C_j^k} \times {a_{k,i}^{F}}^{(n) S_i C_j^{\bar{k}}} \\ &\times (1 - {a_{k,i}^{T}}^{(n)} - {a_{k,i}^{F}}^{(n)})^{(1 - S_i C_j^k - S_i C_j^{\bar{k}})} \Big\} \end{aligned} \tag{9}$$

where $Z_k(n, j)$ is the conditional probability of the claim C_j to have value k given the observation matrix related to the j^{th} claim and current estimate of θ. X_j represents the j^{th} column of the observed SC matrix (i.e., observations of the j^{th} claim from all participants). $A_k(n, j)$ represents the conditional probability regarding observations about the j^{th} claim and current estimation of the parameter θ given the j^{th} claim is of value k.

The Maximization step (M-Step) is given by (5). We choose θ^* (i.e., $({a^T_{k,1}}^*, ...{a^T_{k,M}}^*$; ${a^F_{k,1}}^*, ...{a^F_{k,M}}^*$; d^*) $k = 1, 2, ...K$) that maximizes the $Q\left(\theta|\theta^{(n)}\right)$ function in each iteration to be the $\theta^{(n+1)}$ of the next iteration.

To get θ^* that maximizes $Q\left(\theta|\theta^{(n)}\right)$, we set the derivatives $\frac{\partial Q}{\partial a^T_{k,i}} = 0$, $\frac{\partial Q}{\partial a^F_{k,i}} = 0$ and $\frac{\partial Q}{\partial d_k} = 0$.

Solving the above equations, we can get expressions of the optimal ${a^T_{k,i}}^*$, ${a^F_{k,i}}^*$ and d^*_k:

$$
\begin{aligned}
{a^T_{k,i}}^{(n+1)} &= {a^T_{k,i}}^* = \frac{\sum_{j \in SJ^k_i} Z_k(n,j)}{\sum_{j=1}^{N} Z_k(n,j)} \\
{a^F_{k,i}}^{(n+1)} &= {a^F_{k,i}}^* = \frac{\sum_{j \in SJ^{\bar{k}}_i} Z_k(n,j)}{\sum_{j=1}^{N} Z_k(n,j)} \\
{d_k}^{(n+1)} &= {d_k}^* = \frac{\sum_{j=1}^{N} Z_k(n,j)}{N}
\end{aligned}
\tag{10}
$$

where N is the total number of claims in the conflicting observation matrix. SJ^k_i are the sets of claims the participant S_i actually observes to have value k and $SJ^{\bar{k}}_i$ are the ones S_i observes to have value other than k in the conflicting observation matrix (i.e, SC). $Z_k(n,j)$ is defined in (8). For details of deriving the above solution, please refer to the appendix in Section 8. Note that the case where the value of the claim is binary (i.e., $K = 2$) can be considered as a special case of the algorithm derived in this section. The E-step and M-step of the algorithm for binary claims can be written as in (11) and (12) respectively:

$$
\begin{aligned}
&Q\left(\theta|\theta^{(n)}\right) = \\
&\sum_{j=1}^{N} \Bigg\{ Z_1(n,j) \times \Big[\sum_{i=1}^{M} \Big(S_i C^1_j \log a^T_{1,i} + S_i C^2_j \log a^F_{1,i} \\
&+ (1 - S_i C^1_j - S_i C^2_j) \log(1 - a^T_{1,i} - a^F_{1,i}) + \log d_1 \Big) \Big] \\
&+ (1 - Z_1(n,j)) \times \Big[\sum_{i=1}^{M} \Big(S_i C^2_j \log a^T_{2,i} + S_i C^1_j \log a^F_{2,i} \\
&+ (1 - S_i C^1_j - S_i C^2_j) \log(1 - a^T_{2,i} - a^F_{2,i}) + \log(1 - d_1) \Big) \Big] \Bigg\}
\end{aligned}
\tag{11}
$$

where $S_i C^k_j = 1, k = 1, 2$ when S_i reports C_j to have value k and 0 otherwise. Note that $Z_2(n,j) = 1 - Z_1(n,j)$ and $d_2 = 1 - d_1$ for the binary case.

$$
\begin{aligned}
{a_{1,i}^T}^{(n+1)} &= {a_{1,i}^T}^* = \frac{\sum_{j \in SJ_i^1} Z_1(n,j)}{\sum_{j=1}^N Z_1(n,j)} \\
{a_{1,i}^F}^{(n+1)} &= {a_{1,i}^F}^* = \frac{\sum_{j \in SJ_i^2} Z_1(n,j)}{\sum_{j=1}^N Z_1(n,j)} \\
{a_{2,i}^T}^{(n+1)} &= {a_{2,i}^T}^* = \frac{K_i^1 - \sum_{j \in SJ_i^1} Z_1(n,j)}{N - \sum_{j=1}^N Z_1(n,j)} \\
{a_{2,i}^F}^{(n+1)} &= {a_{2,i}^F}^* = \frac{K_i^2 - \sum_{j \in SJ_i^2} Z_1(n,j)}{N - \sum_{j=1}^N Z_1(n,j)} \\
{d_1}^{(n+1)} &= {d_1}^* = \frac{\sum_{j=1}^N Z_1(n,j)}{N}
\end{aligned} \tag{12}
$$

where SJ_i^1 and SJ_i^2 are the sets of claims S_i reports to have one of the binary values respectively and K_i^1 and K_i^2 are the number of claims in the above two sets.

This completes the mathematical development. We summarize the EM algorithm to handle conflicting observations in the next subsection.

3.3. The Conflict EM Algorithm

We call the EM scheme derived above to handle conflicting observations the EM-Conflict algorithm. The input to the EM-Conflict algorithm is the conflicting observation matrix (i.e., SC) and the output is the maximum likelihood estimation of participant reliability and corresponding judgment on the correctness of claims in the context of conflicting observations. The E-step and M-step of the conflict EM algorithm reduce to simply calculating (8) and (10) iteratively until they converge. The convergence analysis has been done for EM scheme and it is beyond the scope of this paper [Wu 1983]. In practice, we can run the algorithm until the difference of estimation parameter between consecutive iterations becomes insignificant. We can then decide the value of claim C_j as the one that has the highest $Z_k(n,j)$ value for $k = 1, 2, ...K$. In the special case where the claim is binary, C_j is true if $Z_k(n,j) \geq 0.5$ and false otherwise. At the same time, we can also compute the maximum likelihood estimation on participant reliability from the converged values of $\theta^{(n)}$ based on (2). We summarize the resulting algorithm as shown in Algorithm 1.

4. TRUTH DISCOVERY MODEL FOR SOCIAL SENSING WITH CORROBORATING OBSERVATIONS

4.1. The Truth Discovery Model for Corroborating Observations

In the previous section, we proposed the truth discovery model and the EM-Conflict algorithm for conflicting observations of social sensing applications. However, there are some social sensing applications where observations are only *corroborating*. For example, a group of drivers might join a campaign to report freeway locations in need of repair. In this application, only locations in need of repair are reported and locations of normal condition are generally not reported. Hence, in this section we assume, for the model to handle corroborating observations, that the "normal" state of the claim is negative (e.g., no potholes on streets). Hence, participants report only when a positive value is encountered and their observations are corroborating with each other. Based on this assumption, the model for corroborating observations can be treated as a spe-

ALGORITHM 1: Conflict Expectation Maximization Algorithm for Conflicting Observations

Initialize θ with random values between 0 and 1
while $\theta^{(n)}$ does not converge **do**
  for $j = 1 : N$ **do**
    for $k = 1 : K$ **do**
      compute $Z_k(n, j)$ based on (8)
    end for
  end for
  $\theta^{(n+1)} = \theta^{(n)}$
  for $k = 1 : K$ **do**
    for $i = 1 : M$ **do**
      compute $a_{k,i}^{T\ (n+1)}, a_{k,i}^{F\ (n+1)}$ and $d_k^{(n+1)}$ based on (10)
      update $a_{k,i}^{T\ (n)}, a_{k,i}^{F\ (n)}, d_k^{(n)}$ with $a_{k,i}^{T\ (n+1)}, a_{k,i}^{F\ (n+1)}$ and $d_k^{(n+1)}$ in $\theta^{(n+1)}$
    end for
  end for
  $n = n + 1$
end while
Let $Z_{k,j}^c$ = converged value of $Z_k(n, j)$
Let θ^c = converged value of $\theta^{(n)}$
for $j = 1 : N$ **do**
  $max = 0; k^* = 0$
  for $k = 1 : K$ **do**
    if $Z_{k,j}^c \geq max$ **then**
      $max = Z_{k,j}^c$ and $k^* = k$
    end if
  end for
  Claim C_j is of value k^*
end for
for $i = 1 : M$ **do**
  calculate t_i^* from θ^c based on (2).
end for
Return the computed maximum likelihood estimation on source reliability t_i^* and corresponding judgment on the true value k^* of claim C_j.

cial case of the model we discussed in Section 3.1 for conflicting observations when conflicting observations are never asserted on the same claim and the claim is binary.

Since participants only report positive observations of the claim, we simplify some of our notations used in the previous section for the model with corroborating observations. In particular, S_iC_j denotes an observation reported by participant S_i claiming that C_j is true (e.g., that a given street is in disrepair). Let the probability that participant S_i makes an observation be s_i. The same as before, 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*.

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

$$a_i = P(S_iC_j|C_j^t)$$
$$b_i = P(S_iC_j|C_j^f) \tag{13}$$

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

$$a_i = \frac{t_i \times s_i}{d}$$
$$b_i = \frac{(1 - t_i) \times s_i}{1 - d} \tag{14}$$

where the background bias d is the overall probability that a randomly chosen claim 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. Note also that the probability that a participant makes an observation (i.e., s_i) is proportional to the number of claims observed by the participant over the total number of claims observed by all participants, which can be easily computed from the observation matrix.

The only input to our algorithm is the social sensing topology with corroborating observations 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 *corroborating observation matrix*.

4.2. Expectation Maximization for Corroborating Observations

The likelihood function $L'(\theta; X, Z)$ of the above model for corroborating claims is given by:

$$\begin{aligned} &L'(\theta; X, Z) = p(X, Z|\theta) \\ &= \prod_{j=1}^{N} \left\{ \prod_{i=1}^{M} a_i^{S_iC_j} (1 - a_i)^{(1 - S_iC_j)} \times d \times z_j \right. \\ &\left. + \prod_{i=1}^{M} b_i^{S_iC_j} (1 - b_i)^{(1 - S_iC_j)} \times (1 - d) \times (1 - z_j) \right\} \end{aligned} \tag{15}$$

where $\theta = (a_1, a_2, ...a_M; b_1, b_2, ...b_M; d)$ is the simplified estimation parameter of the model for corroborating observations. a_i and b_i are the conditional probabilities defined in (13), and $z_j = 1$ if claim C_j is true and 0 otherwise. d is the background bias that a randomly chosen claim is true. $S_iC_j = 1$ when participant S_i reports that C_j is true, and $S_iC_j = 0$ otherwise. The values of S_iC_j are statistically independent over the M participants and N claims. The likelihood function above describes the likelihood to have current observed data X and hidden variable Z given the estimation parameter θ we defined.

Given the above formulation, substitution of the likelihood function defined in (15) into the definition of Q given by (4) leads to the E-step:

$$\begin{aligned} &Q\left(\theta|\theta^{(n)}\right) \\ &= \sum_{j=1}^{N} \left\{ Z'(n,j) \times \left[\sum_{i=1}^{M} (S_iC_j \log a_i + (1 - S_iC_j) \log(1 - a_i) + \log d) \right] \right. \\ &\left. + (1 - Z'(n,j)) \times \left[\sum_{i=1}^{M} (S_iC_j \log b_i + (1 - S_iC_j) \log(1 - b_i) + \log(1 - d)) \right] \right\} \end{aligned} \tag{16}$$

where $Z'(n, j)$ is given by:

$$
\begin{aligned}
Z'(n,j) &= p(z_j = 1|X_j, \theta^{(n)}) \\
&= \frac{p(z_j = 1; X_j, \theta^{(n)})}{p(X_j, \theta^{(n)})} \\
&= \frac{p(X_j, \theta^{(n)}|z_j = 1)p(z_j = 1)}{p(X_j, \theta^{(n)}|z_j = 1)p(z_j = 1) + p(X_j, \theta^{(n)}|z_j = 0)p(z_j = 0)} \\
&= \frac{A'(n,j) \times d^{(n)}}{A'(n,j) \times d^{(n)} + B'(n,j) \times (1 - d^{(n)})}
\end{aligned} \tag{17}
$$

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

$$
\begin{aligned}
A'(n,j) &= p(X_j, \theta^{(n)}|z_j = 1) \\
&= \prod_{i=1}^{M} a_i^{(n)S_iC_j}(1 - a_i^{(n)})^{(1-S_iC_j)} \\
B'(n,j) &= p(X_j, \theta^{(n)}|z_j = 0) \\
&= \prod_{i=1}^{M} b_i^{(n)S_iC_j}(1 - b_i^{(n)})^{(1-S_iC_j)}
\end{aligned} \tag{18}
$$

The Maximization step (M-step) is given by Equation (5). We choose θ^* (i.e., $(a_1^*, a_2^*, ...a_M^*; b_1^*, b_2^*, ...b_M^*; d^*)$) that maximizes the $Q\left(\theta|\theta^{(n)}\right)$ function in each iteration to be the $\theta^{(n+1)}$ of the next iteration.

To get θ^* that maximizes $Q\left(\theta|\theta^{(n)}\right)$, we set the derivatives $\frac{\partial Q}{\partial a_i} = 0$, $\frac{\partial Q}{\partial b_i} = 0$, $\frac{\partial Q}{\partial d} = 0$. Solving the above equations, we can get expressions of the optimal a_i^*, b_i^* and d^*:

$$
\begin{aligned}
a_i^{(n+1)} = a_i^* &= \frac{\sum_{j \in SJ_i'} Z'(n,j)}{\sum_{j=1}^{N} Z'(n,j)} \\
b_i^{(n+1)} = b_i^* &= \frac{K_i - \sum_{j \in SJ_i'} Z'(n,j)}{N - \sum_{j=1}^{N} Z'(n,j)} \\
d_i^{(n+1)} = d_i^* &= \frac{\sum_{j=1}^{N} Z'(n,j)}{N}
\end{aligned} \tag{19}
$$

where K_i is the number of claims observed by participant S_i and N is the total number of claims in the observation matrix. SJ_i' is the set of claims the participant S_i actually observes in the observation matrix SC'. $Z'(n,j)$, defined in (17), is the probability the j^{th} claim is true given the observed data and current estimation of θ. For details of deriving the above solution, please refer to the appendix in Section 8.

Given the above, The E-step and M-step of EM optimization reduce to simply calculating (17) and (19) iteratively until they converge. Since the claim is binary, C_j is true if $Z'(n,j) \geq 0.5$ and false otherwise. At the same time, we can also compute the maximum likelihood estimation of participant reliability t_i^* from the converged values of $a_i^{(n)}$, $b_i^{(n)}$ and $d^{(n)}$ based on their relationship given by (14). We summarize the resulting algorithm in the subsection below.

4.3. The Regular EM Algorithm

ALGORITHM 2: Regular Expectation Maximization Algorithm

Initialize θ with random values between 0 and 1
while $\theta^{(n)}$ does not converge **do**
  for $j = 1 : N$ **do**
    compute $Z'(n, j)$ based on (17)
  end for
  $\theta^{(n+1)} = \theta^{(n)}$
  for $i = 1 : M$ **do**
    compute $a_i^{(n+1)}, b_i^{(n+1)}, d^{(n+1)}$ based on (19)
    update $a_i^{(n)}, b_i^{(n)}, d^{(n)}$ with $a_i^{(n+1)}, b_i^{(n+1)}, d^{(n+1)}$ in $\theta^{(n+1)}$
  end for
  $n = n + 1$
end while
Let Z_j^c = converged value of $Z'(n, j)$
Let a_i^c = converged value of $a_i^{(n)}$; b_i^c = converged value of $b_i^{(n)}$; d^c = converged value of $d^{(n)}$
for $j = 1 : N$ **do**
  if $Z_j^c \geq 0.5$ **then**
    C_j is true
  else
    C_j is false
  end if
end for
for $i = 1 : M$ **do**
  calculate t_i^* from a_i^c, b_i^c and d^c based on (14)
end for
Return the computed maximum likelihood estimation on source reliability t_i^* and corresponding judgment on the correctness of claim C_j.

We call the EM scheme derived above the regular EM algorithm as it handles only corroborating observations of a claim from different participants. The input to the regular EM algorithm is the corroborating observation matrix SC' from social sensing data, and the output is the maximum likelihood estimation of participant reliability and correctness judgment of claims. The pseudocode of regular EM is shown in Algorithm 2.

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[Cramer 1946]. 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 at this bound and hence reach the maximum likelihood estimation [Wang et al. 2012b]. This observation makes it possible to quantify estimation accuracy, or confidence in results generated from our scheme, using the Cramer-Rao lower bound [Wang et al. 2013b].

5. EVALUATION

In this section, we carry out experiments to evaluate the performance of the proposed conflict EM scheme (i.e, EM-Conflict) in terms of estimation accuracy of the probability that a participant is right or a claim is true compared to the regular EM scheme described in Section 4 and other state-of-art solutions for conflicting observations. Those baselines include Sums [Kleinberg 1999], Average-Log [Pasternack and Roth 2010],

TruthFinder [Yin et al. 2008] and the simple voting scheme. We begin by considering algorithm performance for different abstract observation matrices (i.e., SC), then apply it to both an emulated participatory sensing scenario and a real world social sensing application. We show that the conflict EM algorithm outperforms the regular EM algorithm and the state of the art baselines.

5.1. A Simulation Study

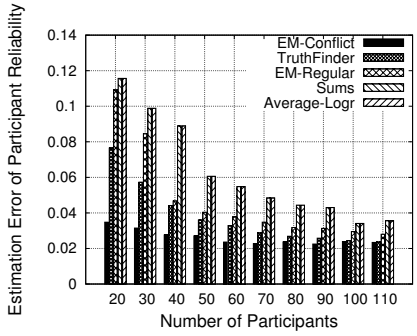
We built a simulator in Matlab 7.10.0 and evaluated the performance of the proposed EM scheme for conflicting observations compared to the regular EM scheme and several state-of-art techniques in literature. The simulator generates a random number of participants and claims. A random probability P_i 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 from participant S_i has a probability t_i of being matched to the correctness of the claim (i.e., reporting a variable to be the same as its ground truth) and a probability $1 - t_i$ of being mismatched. Note that participants can report conflicting observations for the same claim in this scenario. For simplicity, we assume claims to be binary.

We applied both the conflict EM scheme derived in section 3 and the regular EM scheme derived in section 4 to the sensing topology with conflicting observations and showed that the conflict EM scheme outperformed the regular EM scheme and other state-of-art baselines. Note that for the regular EM scheme, we adapted it in a similar way as fact-finders to handle the conflicting observations of the same claim [Pasternack and Roth 2010]. Specifically, it takes conflicting observations of the same claim as two independent observations and pick the one with higher probability to believe after the algorithm terminates. Reported results are averaged over 100 experiments.

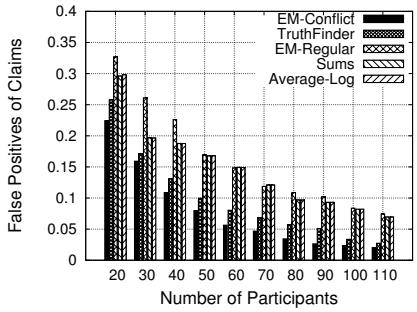
In the first experiment, we compare the estimation accuracy of the conflict EM scheme and baselines (including regular EM scheme) by varying the number of participants in the network. The number of reported claims was fixed at 2000, of which 1000 claims were reported correctly and 1000 were misreported. The average number of observations per participant was set to 200. The number of participants was varied from 20 to 110. Results are shown in Figure 1. Observe that the conflict EM scheme has both smaller estimation error on participant reliability and less false positives/negatives on claims among all schemes under comparison. Note also that the performance gain of the conflict EM scheme is large when the number of participants is small.

The second experiment compares the conflict EM scheme with baselines when the average number of observations per participant changes. As before, we fixed the number of correctly and incorrectly reported claims to be 1000 respectively. The number of participants was fixed at 50. We vary the average number of observations per participant from 100 to 1000. The results are shown in Figure 2. Observe that the EM scheme outperforms the regular EM scheme and other baselines in terms of both participant reliability estimation accuracy and false positives/negatives of claims as the average number of observations per participant changes. The performance gain of the conflict EM scheme is higher when the average number of observations per participant is low.

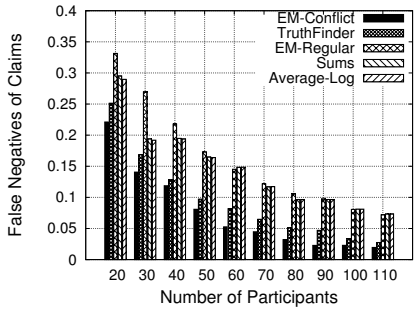
The third experiment examines the effect of changing the percentage of correct claims on the estimation accuracy of all schemes. We varied the ratio of the number of correctly reported claims to the total number of claims from 0.1 to 0.6, while fixing the total number of such claims to 2000. The number of participants was fixed to 50 and the average number of observations reported by a participant was set to 200. Reported results are shown in Figure 3. We observe that the conflict EM scheme has less error



(a) Participant Reliability Estimation Accuracy



(b) Claim Estimation: False Positives

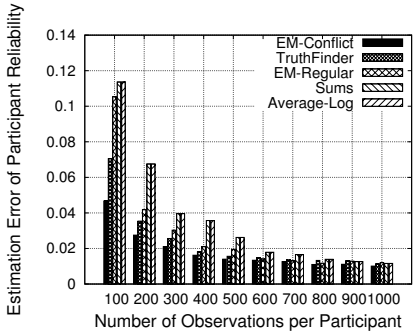


(c) Claim Estimation: False Negatives

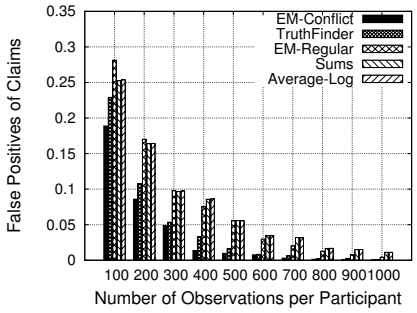
Fig. 1. Estimation Accuracy versus Number of Participants for Conflicting Observations

in both participant reliability estimation and false positives/negatives on claims under different mix of correct and false claims.

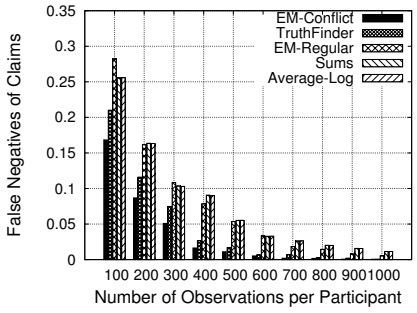
Finally, we carried out the fourth experiment to evaluate the performance of the conflict EM scheme and other schemes when the offset of the initial estimation on the background bias d varies. The offset is defined as the difference between the initial estimation on d and its ground-truth. We fixed the number of correctly and incorrectly



(a) Participant Reliability Estimation Accuracy



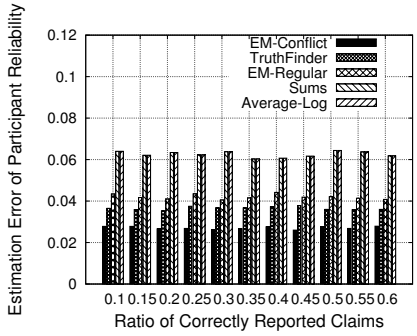
(b) Claim Estimation: False Positives



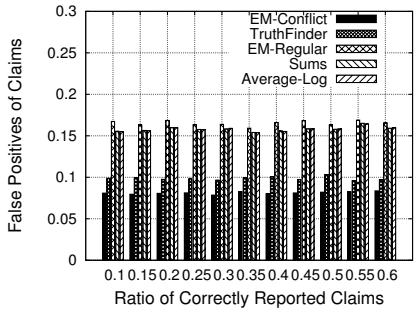
(c) Claim Estimation: False Negatives

Fig. 2. Estimation Accuracy versus Average Number of Observations per Participant for Conflicting Observations

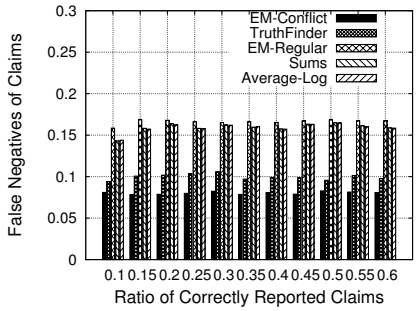
reported claims to 1000 respectively (i.e., $d = 0.5$). We varied the absolute value of the initial estimate offset on d from 0 to 0.45. The number of participants was fixed at 50 and the average number of observations per participant was set to 200. Results are averaged over both positive and negative offsets of the same absolute value. Figure 4 shows the results. We observe that the performance of the conflict EM scheme is bet-



(a) Participant Reliability Estimation Accuracy



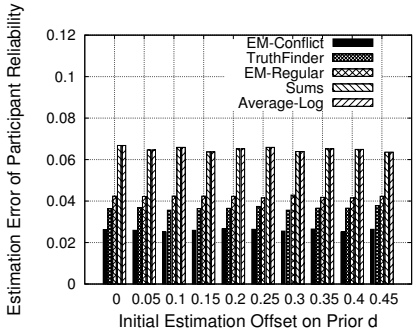
(b) Claim Estimation: False Positives



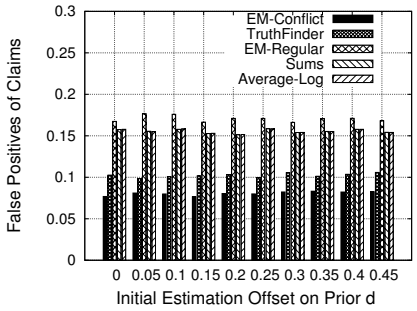
(c) Claim Estimation: False Negatives

Fig. 3. Estimation Accuracy versus Ratio of Correctly Reported Claims for Conflicting Observations

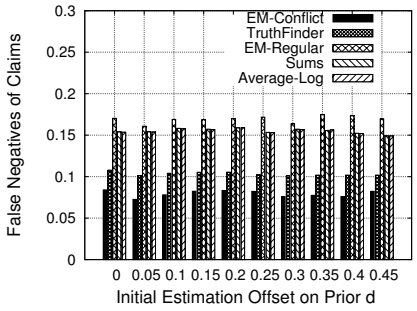
ter than other baselines in terms of both participant reliability estimation and false positives/negatives on claims when the initial estimate offset on d changes. We also observe the performance of all schemes are relatively stable when offset on d increases. The reason is the schemes for conflicting observations mainly depend on the mutual exclusive property of the reports (rather than correct estimation on prior d) to decide the correctness of claims.



(a) Participant Reliability Estimation Accuracy



(b) Claim Estimation: False Positives



(c) Claim Estimation: False Negatives

Fig. 4. Estimation Accuracy versus Initial Estimation Offset on Prior d for Conflicting Observations

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 locations and report *whether or not* litters exist in the tagged locations. However, their reports may potentially be conflicting and are not reliable, 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 conflict EM scheme to several baselines including the best performed fact-finder scheme in this scenario, the regular EM scheme adapted for conflicting observations and voting, where locations are simply ranked by the number of times people report them.

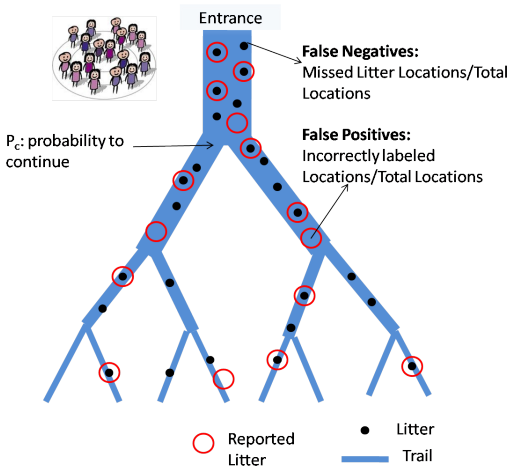
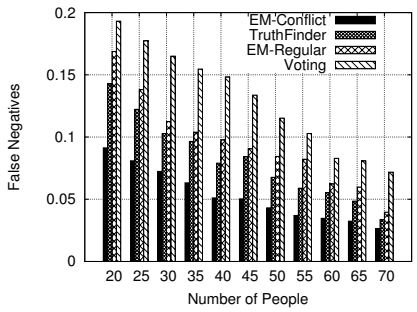


Fig. 5. A Simplified Trail Map of Geotagging Application

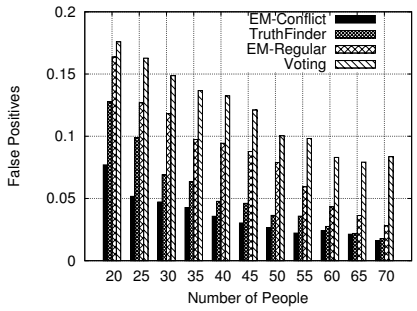
We created a simplified trail map of a park, represented by a binary tree as shown in Figure 5. The entrance of the park (e.g., where parking areas are usually located) is the root of the tree. Internal nodes of the tree represent forking of different trails. We assume trails are quantized into discretely labeled locations (e.g., numbered distance markers). In our 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 and vice versa).

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 0.2 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 75%. We vary the number of participants visiting the park from 20 to 70. The corresponding estimation results of different schemes are shown in Figure 6. 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 conflict EM scheme outperforms others in terms of both false negatives and false positivies, which means the conflict EM scheme can find more pieces of litter than other schemes while keeping the falsely reported locations less. Generally, voting performs the worst in accuracy because it simply counts the number of reports claiming about each location but ignores the reliability of individuals who make them.



(a) False Negatives (missed/total litter)

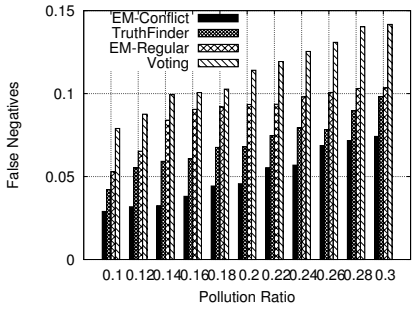


(b) False Positives (false/total locations)

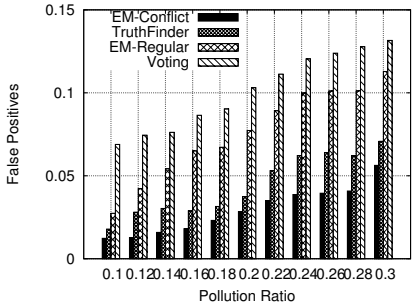
Fig. 6. Litter Geotagging Accuracy versus Number of People Visiting the Park

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 50. We vary the pollution ratio of the park from 0.1 to 0.3. The estimation results of different schemes are shown in Figure 7. 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 conflict EM scheme continues to find more actual litter locations and report less falsely labeled locations compared to other baselines.



(a) False Negatives (missed/total litter)



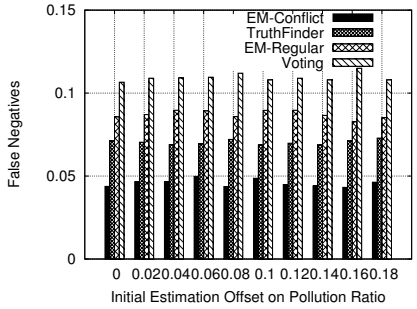
(b) False Positives (false/total locations)

Fig. 7. Litter Geotagging Accuracy versus Pollution Ratio of the Park

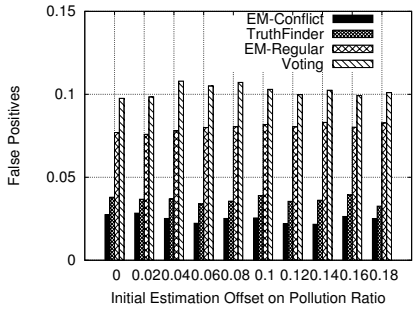
In the third experiment, we evaluate the effect of the initial estimation offset of the pollution ratio on the performance of different schemes. The pollution ratio is fixed at 0.2 and the number of individuals visiting the park is set to be 50. We vary the absolute value of initial estimation offset of the pollution ratio from 0 to 0.18. Results are averaged over both positive and negative offsets of the same absolute value. The estimation results of different schemes are shown in Figure 8. Observe that the conflict EM scheme finds more actual litter locations and reports less falsely labeled locations than other baselines throughout all initial estimation offsets of pollution ratio simulated.

5.3. A Real World Application

In this section, we evaluate the performance of the proposed conflict EM scheme compared to start-of-art baselines (including the regular EM scheme) through a real world application, finding free parking lots on University of Illinois at Urbana Champaign (UIUC) campus. "Free parking lots"' refer to the parking lots that are free of charge after 5pm on weekday as well as weekends. The goal was to see if our scheme can find the free parking lots most accurately compared to other state-of-art baselines. Specifically,



(a) False Negatives (missed/total litter)



(b) False Positives (false/total locations)

Fig. 8. Litter Geotagging Accuracy versus Initial Estimation Offset on Pollution Ratio of Park

we selected 106 parking lots of our interests around the campus and asked volunteers to mark them as either "Free"' or "Not Free"'. Participants mark those parking lots they have been to or are familiar with. We note that there are actually various types of parking lots on campus: enforced parking lots with time limits, parking meters, permit parking, street parking, and etc. Different parking lots have different regulations for free parking. Moreover, instructions and permit signs sometimes read similar and are easy to miss. Hence, people are prone to generate both false positives and false negatives in their reports. For evaluation purpose, we went to those selected parking lots and manually collected the ground truth.

In the experiment, 30 participants were invited to offer their marks on the 106 parking lots (46 of which are indeed free). There were 901 marks collected from participants in total. We then generated the observation matrix by taking the participants as sources and different parking lots as claims. The free parking lots are taken as the true claims while the non-free ones are taken as the false claims. The corresponding element S_iC_j is set according to the marks each participant placed on those parking lots. We applied the conflict EM scheme discussed in Section 3, other state-of-art baselines (including the regular EM scheme adapted for conflicting claims) as well as the simple voting scheme to the data we collected. We then compared the false positives and false negatives of different schemes in identifying the free parking lots among all places selected. The result is shown in Table I. We observe that the conflict EM scheme designed to handle conflicting observations (i.e., EM-Conflict) achieved the least false

Table I. Accuracy of Finding Free Parking Lots on Campus

Schemes	False Positives	False Negatives
EM-Conflict	**6.67%**	**10.87%**
EM-Regular	11.67%	17.39%
Average-Log	16.67%	19.57%
Truth-Finder	18.33%	15.22%
Voting	21.67%	23.91%

positives and false negatives among all schemes under comparison. The reason is that the conflict EM scheme modeled the conflicting observations explicitly and used the MLE approach to find the value of each claim that is most consistent with the observations we had.

The above evaluations demonstrate that the new EM scheme for conflicting observations 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 or using some heuristic models [Pasternack and Roth 2010; Yin et al. 2008; Wang et al. 2011a]. However, the conflict EM scheme explicitly models the conflicting observations and makes its inference based on a maximum likelihood hypothesis that is most consistent with the observed sensing data.

6. DISCUSSION AND LIMITATIONS

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 duplicated observations from dependent sources and only keep the original ones. This can be achieved by applying copy detection schemes between sources [Dong et al. 2009; Dong et al. 2010]. Another possible solution is to cluster dependent sources based on some *source-dependency* metric [Berti-Equille et al. 2009]. 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.

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 [Wang et al. 2012a]. 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 parameter of interests 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 EM, would be a better fit [Wang et al. 2013a]. 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 where observations from participants may be *conflicting*. The approach can determine the correctness of reported observations given only the measurements sent without knowing the trustworthiness of

participants. The maximum likelihood 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.

8. APPENDIX

The following derivation demonstrates the details to obtain the results in (10). The derivation that maximizes the $Q\left(\theta|\theta^{(t)}\right)$ in the M-step in Section 3.2 yields:

$$\sum_{j=1}^{N} Z_k(n,j)\left[\frac{S_iC_j^k}{a_{k,i}^{T\ *}} - \frac{(1-S_iC_j^k - S_iC_j^{\bar{k}})}{1-a_{k,i}^{T\ *} - a_{k,i}^{F\ *}}\right] = 0$$

$$\sum_{j=1}^{N} Z_k(n,j)\left[\frac{S_iC_j^{\bar{k}}}{a_{k,i}^{F\ *}} - \frac{(1-S_iC_j^k - S_iC_j^{\bar{k}})}{1-a_{k,i}^{T\ *} - a_{k,i}^{F\ *}}\right] = 0 \quad k = 1,2,..K \tag{20}$$

$$\sum_{j=1}^{N}\left[Z_k(n,j)\frac{1}{d_k^*} - Z_K(n,j))\frac{1}{1-\sum_{i=1}^{K-1} d_i^*}\right] = 0 \quad k = 1,2,..K-1 \tag{21}$$

As we defined earlier, SJ_i^k and $SJ_i^{\bar{k}}$ represent the sets of claims the participant S_i actually reports as value k and value other than k respectively in the conflicting observation matrix (i.e, SC). Let us also define $\bar{SJ}_i$ as the set of claims participant S_i does not report in the conflicting observation matrix. Thus, (20) can be rewritten as:

$$\sum_{j\in SJ_i^k} Z_k(n,j)\frac{1}{a_{k,i}^{T\ *}} - \sum_{j\in \bar{SJ}_i} Z_k(n,j)\frac{1}{1-a_i^{T*} - a_{k,i}^{F\ *}} = 0$$

$$\sum_{j\in SJ_i^{\bar{k}}} Z_k(n,j)\frac{1}{a_{k,i}^{F\ *}} - \sum_{j\in \bar{SJ}_i} Z_k(n,j)\frac{1}{1-a_{k,i}^{T\ *} - a_{k,i}^{F\ *}} = 0 \tag{22}$$

Solving the above equations, we can obtain the expressions of the optimal $a_{k,i}^{T\ *}$, $a_{k,i}^{F\ *}$ and d_k^* as shown in (10).

Similarly, the following derivation demonstrates the details to obtain the results in (19). The derivation that maximizes the $Q\left(\theta|\theta^{(n)}\right)$ in the M-step in Section 4.2 yields:

$$\sum_{j=1}^{N}\left[Z'(n,j)(S_iC_j\frac{1}{a_i^*} - (1-S_iC_j)\frac{1}{1-a_i^*})\right] = 0$$

$$\sum_{j=1}^{N}\left[(1-Z'(n,j))(S_iC_j\frac{1}{b_i^*} - (1-S_iC_j)\frac{1}{1-b_i^*})\right] = 0$$

$$\sum_{j=1}^{N}\left[Z'(n,j)M\frac{1}{d^*} - (1-Z'(n,j))M\frac{1}{1-d^*})\right] = 0 \tag{23}$$

As we defined before, SJ_i' is the set of claims the participant S_i actually reports in the observation matrix SC, and $\bar{SJ}_i'$ as the set of claims participant S_i does not observe. Thus, (23) can be rewritten as:

$$\sum_{j \in SJ'_i} Z(n,j)\frac{1}{a_i^*} - \sum_{j \in S\bar{J}'_i} Z(n,j)\frac{1}{1-a_i^*} = 0$$

$$\sum_{j \in SJ'_i} (1-Z(n,j))\frac{1}{b_i^*} - \sum_{j \in S\bar{J}'_i} (1-Z(n,j))\frac{1}{1-b_i^*} = 0$$

$$\sum_{j=1}^{N} \left[Z(n,j)\frac{1}{d^*} - (1-Z(n,j))\frac{1}{1-d^*}) \right] = 0 \tag{24}$$

Solving the above equations, we can get expressions of the optimal a_i^*, b_i^* and d^* as shown in (19).

REFERENCES

ABDELZAHER, T. ET AL. 2007. Mobiscopes for human spaces. *IEEE Pervasive Computing 6,* 2, 20–29.

ADOMAVICIUS, G. AND TUZHILIN, A. 2005. 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.

AHMADI, H., ABDELZAHER, T., HAN, J., PHAM, N., AND GANTI, R. 2011. The sparse regression cube: A reliable modeling technique for open cyber-physical systems. In *Proc. 2nd International Conference on Cyber-Physical Systems (ICCPS'11)*.

AHMADI, H., PHAM, N., GANTI, R., ABDELZAHER, T., NATH, S., AND HAN, J. 2010. Privacy-aware regression modeling of participatory sensing data. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*. SenSys '10. ACM, New York, NY, USA, 99–112.

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

BALAN, R. K., KHOA, N. X., AND JIANG, L. 2011. Real-time trip information service for a large taxi fleet. In *Proceedings of the ninth international conference on Mobile systems, applications, and services (MobiSys'11)*.

BECKER, S. I. R., CCERES, R., ROWLAND, M. M. J., VARSHAVSKY, A., AND WILLINGER, W. 2012. Human mobility modeling at metropolitan scales. In *Proceedings of the tenth international conference on Mobile systems, applications, and services (MobiSys'12)*.

BERTI-EQUILLE, L., SARMA, A. D., DONG, X., MARIAN, A., AND SRIVASTAVA, D. 2009. Sailing the information ocean with awareness of currents: Discovery and application of source dependence. In *CIDR'09*.

BRIN, S. AND PAGE, L. 1998. The anatomy of a large-scale hypertextual web search engine. In *7th international conference on World Wide Web (WWW'07)*. 107–117.

CRAMER, H. 1946. *Mathematical Methods of Statistics*. Princeton Univ. Press.

DEKEL, O. AND SHAMIR, O. 2009. Vox populi: Collecting high-quality labels from a crowd. In *In Proceedings of the 22nd Annual Conference on Learning Theory*.

DELRE, S. A., JAGER, W., AND JANSSEN, M. A. 2007. Diffusion dynamics in small-world networks with heterogeneous consumers. *Comput. Math. Organ. Theory 13*, 185–202.

DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B 39,* 1, 1–38.

DONG, X., BERTI-EQUILLE, L., HU, Y., AND SRIVASTAVA, D. 2010. Global detection of complex copying relationships between sources. *PVLDB 3,* 1, 1358–1369.

DONG, X., BERTI-EQUILLE, L., AND SRIVASTAVA, D. 2009. Truth discovery and copying detection in a dynamic world. *VLDB 2,* 1, 562–573.

DOUCET, A., DE FREITAS, N., AND GORDON, N., Eds. 2001. *Sequential Monte Carlo methods in practice*.

DUDA, R. O., HART, P. E., AND STORK, D. G. 2001. *Pattern Classification (2nd Edition)* 2 Ed. Wiley-Interscience.

EISENMAN, S. B. ET AL. 2007. The bikenet mobile sensing system for cyclist experience mapping. In *SenSys'07*.

GALLAND, A., ABITEBOUL, S., MARIAN, A., AND SENELLART, P. 2010. Corroborating information from disagreeing views. In *WSDM*. 131–140.

HUANG, J.-H., AMJAD, S., AND MISHRA, S. 2005. CenWits: a sensor-based loosely coupled search and rescue system using witnesses. In *SenSys'05*. 180–191.

HUI, C., GOLDBERG, M. K., MAGDON-ISMAIL, M., AND WALLACE, W. A. 2010. Simulating the diffusion of information: An agent-based modeling approach. *IJATS*, 31–46.

HULL, B. ET AL. 2006. CarTel: a distributed mobile sensor computing system. In *SenSys'06*. 125–138.

INC, U. T. AND STAFF, U. T. I. 1997. *Solving Data Mining Problems Using Pattern Recognition Software with Cdrom* 1st Ed. Prentice Hall PTR, Upper Saddle River, NJ, USA.

J.HAN, M.KAMBER, AND PEI, J. 2011. *Data Mining: Concepts and Techniques, Third Edition*. Morgan Kaufman.

JOHNSON, R. A. AND WICHERN, D. W. 2002. *Applied multivariate statistical analysis*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

KALMAN, R. E. 1960. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME Journal of Basic Engineering* 82 (Series D), 35–45.

KLEINBERG, J. M. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM 46,* 5, 604–632.

MUSTAPHA, N., JALALI, M., AND JALALI, M. 2009. Expectation maximization clustering algorithm for user modeling in web usage mining systems. *European Journal of Scientific Research 32,* 4, 467–476.

NATH, S. 2012. Ace: Exploiting correlation for energy-efficient and continuous context sensing. In *Proceedings of the tenth international conference on Mobile systems, applications, and services (MobiSys'12)*.

PARK, T., LEE, J., HWANG, I., YOO, C., NACHMAN, L., AND SONG, J. 2011. E-gesture: a collaborative architecture for energy-efficient gesture recognition with hand-worn sensor and mobile devices. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*. SenSys '11. ACM, New York, NY, USA, 260–273.

PASTERNACK, J. AND ROTH, D. 2010. Knowing what to believe (when you already know something). In *International Conference on Computational Linguistics (COLING)*.

PHAM, N., GANTI, R. K., UDDIN, Y. S., NATH, S., AND ABDELZAHER, T. 2010. Privacy-preserving reconstruction of multidimensional data maps in vehicular participatory sensing.

POMERANTZ, D. AND DUDEK, G. 2009. 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. Springer-Verlag, Berlin, Heidelberg, 98–109.

REDDY, S., ESTRIN, D., AND SRIVASTAVA, M. 2010a. Recruitment framework for participatory sensing data collections. In *Proceedings of the 8th International Conference on Pervasive Computing*. Springer Berlin Heidelberg, 138–155.

REDDY, S., SHILTON, K., DENISOV, G., CENIZAL, C., ESTRIN, D., AND SRIVASTAVA, M. 2010b. Biketastic: sensing and mapping for better biking. In *Proceedings of the 28th international conference on Human factors in computing systems*. CHI '10. ACM, New York, NY, USA, 1817–1820.

SHENG, V. S., PROVOST, F., AND IPEIROTIS, P. G. 2008. 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. ACM, New York, NY, USA, 614–622.

SUN, Y., YU, Y., AND HAN, J. 2009. Ranking-based clustering of heterogeneous information networks with star network schema. In *15th SIGKDD international conference on Knowledge discovery and data mining (KDD'09)*. 797–806.

WANG, D., ABDELZAHER, T., AHMADI, H., PASTERNACK, J., ROTH, D., GUPTA, M., HAN, J., FATEMIEH, O., AND LE, H. 2011a. On bayesian interpretation of fact-finding in information networks. In *14th International Conference on Information Fusion (Fusion 2011)*.

WANG, D., ABDELZAHER, T., KAPLAN, L., AND AGGARWAL, C. C. 2013a. Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications. In *The 33rd International Conference on Distributed Computing Systems (ICDCS 13)*.

WANG, D., AHMADI, H., ABDELZAHER, T., CHENJI, H., STOLERU, R., AND AGGARWAL, C. 2011b. Optimizing quality-of-information in cost-sensitive sensor data fusion. In *IEEE 7th International Conference on Distributed Computing in Sensor Systems (DCoSS 11)*.

WANG, D., KAPLAN, L., , LE, H. K., AND ABDELZAHER, T. 2012a. On truth discovery in social sensing: A maximum likelihood estimation approach. In *The 11th ACM/IEEE Conference on Information Processing in Sensor Networks (IPSN 12)*.

WANG, D., KAPLAN, L., ABDELZAHER, T., AND AGGARWAL, C. C. 2012b. On scalability and robustness limitations of real and asymptotic confidence bounds in social sensing. In *The 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON 12)*.

WANG, D., KAPLAN, L., ABDELZAHER, T., AND AGGARWAL, C. C. 2013b. On credibility tradeoffs in assured social sensing. *IEEE Journal On Selected Areas in Communication (JSAC) 31,* 6.

WU, C. F. J. 1983. On the convergence properties of the EM algorithm. *The Annals of Statistics 11,* 1, 95–103.

XIE, J., SREENIVASAN, S., KORNISS, G., ZHANG, W., LIM, C., AND SZYMANSKI, B. K. 2011. Social consensus through the influence of committed minorities. *Physical Review E 84,* 1, 011130+.

YIN, X., HAN, J., AND YU, P. S. 2008. Truth discovery with multiple conflicting information providers on the web. *IEEE Trans. on Knowl. and Data Eng. 20*, 796–808.

YIN, X. AND TAN, W. 2011. Semi-supervised truth discovery. In *WWW*. ACM, New York, NY, USA.

ZHAO, B., RUBINSTEIN, B. I. P., GEMMELL, J., AND HAN, J. 2012. A bayesian approach to discovering truth from conflicting sources for data integration. *Proc. VLDB Endow. 5,* 6, 550–561.

ZHOU, P., ZHENG, Y., AND LI, M. 2012. How long to wait?: Predicting bus arrival time with mobile phone based participatory sensing. In *Proceedings of the tenth international conference on Mobile systems, applications, and services (MobiSys'12)*.

ZUBIAGA, A., SPINA, D., AMIG, E., AND GONZALO., J. 2012. Towards real-time summarization of scheduled events from twitter streams. In *"23rd ACM Conference on Hypertext and Social Media (Hypertext 2012)"*.