Confidence-Aware Truth Estimation in Social Sensing Applications

Dong Wang, Chao Huang
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556
dwang5@nd.edu, chuang7@nd.edu

***Abstract*—This paper presents a confidence-aware maximum likelihood estimation framework to solve the truth estimation problem in social sensing applications. Social sensing has emerged as a new paradigm of data collection, where a group of individuals volunteer (or are recruited) to share certain observations or measurements about the physical world. A key challenge in social sensing applications lies in ascertaining the correctness of reported observations from unvetted data sources with unknown reliability. We refer to this problem as *truth estimation*. The prior works have made significant efforts to solve this problem by developing various truth estimation algorithms. However, an important limitation exists: they assumed a data source makes all her/his observations with *the same degree of confidence*, which may not hold in many real-world social sensing applications. In this paper, we develop a new confidence-aware truth estimation scheme that removes this limitation by explicitly considering different degrees of confidence that sources express on the reported data. The new truth estimation scheme solves a maximum likelihood estimation problem to determine both the correctness of collected data and the reliability of data sources. We compare our confidence-aware scheme with the state-of-the-art baselines through both an extensive simulation study and three real world case studies based on Twitter. The evaluation shows that our new scheme outperforms all compared baselines and significantly improves the accuracy of the truth estimation results in social sensing applications.**

***Index Terms*—Social Sensing, Confidence-Aware, Truth Estimation, Data Quality, Maximum Likelihood Estimation, Expectation Maximization**

I. Introduction

This paper presents a confidence-aware maximum likelihood estimation approach to solve the truth estimation problem in social sensing applications. We refer by social sensing to a broad set of sensing applications where individuals volunteer or are recruited to collect and share data about the physical environment. For example, people may report their observations during a hurricane or earthquake and broadcast such real-time information through online social media (e.g., Twitter). Alternatively, they may also download an app on smartphones to report litter locations or potholes on city streets (i.e., geo-tagging). Considering the open data collection paradigm and potentially unreliable nature of unvtted human sources, a key challenge of social sensing applications lies in accurately *ascertaining the correctness of reported data*, which is referred to as *truth estimation problem* in social sensing.

Much of the research in the sensor network community focused on physical sensors. This includes dedicated devices embedded in their environment [20], as well as human-centric sensing devices such as cell-phones and wearables [18]. Recent research proposed challenges with the use of humans as sensors [28]. Clearly, humans differ from traditional physical devices in many respects. Importantly to the truth estimation problem in social sensing, they lack a design specification and a reliability standard, making it hard to define a generic noise model for sources. Hence, many techniques that estimate probability of error for physical sensors do not directly apply in social sensing applications.

Reputation systems have been successfully applied in scenarios where source reliability is the issue [14], [17]. The assumption is that, when sources are observed over time, their reliability is eventually uncovered. Social sensing applications, however, often deal with scenarios, where a new event requires data collection from sources who have not previously participated in other data collection campaigns, or perhaps not been "tested" in the unique circumstances of the current event. For example, a hurricane strikes New Jersey. This is a rare event. We do not know how accurate the individuals who fled the event are at describing damage left behind. No reputation is accumulated for them in such a scenario. Then, an interesting question arises: how do we determine which observations to believe with no prior knowledge on the reliability of data sources?

Truth estimation algorithms have been developed to address the above question [23], [34], [36], [41], [45]. These solutions jointly assess both the reliability of sources and the quality of their observations in an iterative fashion without knowing either of them a priori. However, an important limitation exists in the current solutions: they assumed a data source makes all her/his observations with *the same degree of confidence*. This assumption does not hold in real world social sensing applications where users may express various degrees of confidence on their reported data. For example, in case of a fire disaster, a user may tweet: i) "I saw the building at the corner of Main and Adam street is on fire! Photo taken: http://..." or ii) "RT@Mike: I heard the alarm on Main street, but not sure if there was a fire.". Clearly, the user expresses a higher degree of confidence in the first tweet than the second one.

The main contribution of this paper lies in developing a new confidence-aware truth estimation scheme to explicitly consider *different degrees of confidence* that a source may express on reported observations. In particular, we formulated a confidence-aware truth estimation problem in social sensing as a maximum likelihood estimation problem. We derived an optimal solution, based on expectation maximization, which assigns the true values to data items by observing which source reports what observations as well as the degree of confidence associated with each reported observation. We evaluate the proposed new scheme through both an extensive simulation study and three real world case studies based on Twitter. Evaluation results show that the confidence-aware truth estimation scheme outperforms the state-of-the-art baselines in the literature by significantly improving the accuracy of the truth estimation results. The results of this paper are important because they allow social sensing applications to accurately estimate the data quality and source reliability while taking into account the source confidence on the reported data under a rigorous analytical framework.

The rest of this paper is organized as follows: we review related work in Section II. In Section III, we present the new confidence-aware truth estimation model for social sensing applications. The proposed maximum likelihood estimation framework is discussed in Section IV. Simulation and real world case study results are presented in Section V. We discuss the limitations of current model and future work in Section VI. Finally, we conclude the paper in Section VII.

II. Related Work

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 [1]. The ideas of getting people involved into the loop of the sensing process have been investigated at length in participatory sensing [5] and opportunistic sensing [16] applications. Examples of some early social sensing applications include CenWits [11], CarTel [12], and BikeNet [8]. A recent survey of social sensing [2] covers many sensing challenges in human context such as accommodating energy constraints of mobile sensing devices [19], protecting the privacy of participants [4], and promoting social interactions in different environments [25]. It also suggests that humans can actually *act as sensors* by contributing information through "sensing campaigns" [26] or social data scavenging (e.g., via Twitter and Youtube) [44]. However, the *truth estimation* problem still remains as an open research question in social sensing. This paper develops a confidence-aware truth estimation scheme that allows social sensing applications to accurately quantify the reliability of human sensed data while exploiting the *confidence* sources express with their claims.

Techniques for deriving accurate conclusions from sources whose reliability is unknown are referred to as *fact-finders* in data mining and machine learning literature. A comprehensive survey of fact-finders used in the context of trust analysis of information networks can be found in [9]. One of the early papers on the topic was Hubs and Authorities [15] that uses iterative algorithm to compute both source trustworthiness and claim credibility. Other fact-finding schemes enhanced these basic frameworks by using more refined heuristics [41], incorporating analysis on properties of claims [21] and dependency between sources [7]. More recent works came up with some new fact-finding algorithms to handle the background knowledge [22], quantify the accuracy of source and data credibility [35] and use slot filling systems for multi-dimensional fact-finding [43]. In this paper, we will use insights from the above work to develop a new truth estimation framework that addresses the challenges of unknown source reliability and diverse degrees of source confidence in social sensing applications.

Maximum likelihood estimation (MLE) framework has been used extensively in wireless sensor networks (WSN) for various estimation and data fusion tasks [27], [40]. For example, Sheng et al. developed a MLE method that uses acoustic signal energy measurements from individual sensors to estimate locations of multiple sources [27]. Xiao et al. presented a distributed consensus based MLE approach to compute the unknown parameters of sensory measurements corrupted by Gaussian noise [40]. The MLE framework has also been applied to address clock synchronization [39], target tracking [37], and compressive sensing [6] in WSN. However, the above work primarily focused on the estimation of continuous variables from measurements of physical sensors. In this paper, we focus on estimating a set of binary variables that represent either true or false statements about the physical world from human sensed observations. The MLE problem we solved is actually harder due to the discrete nature of the estimated variables and the inherent complexity of modeling humans as sensors in social sensing.

Finally, our work is also related with a type of information filtering system called recommendation systems [13]. Expectation Maximization (EM) has been used as an optimization approach for both collaborative filtering [31] and content based recommendation systems [24]. For example, Wang et al. developed a collaborative filtering based system using the EM approach to recommend scientific articles to users of an online community [31]. Pomerantz et al. proposed a content-based system using EM to explore the contextual information to recommend movies [24]. However, the truth estimation in social sensing studies a different problem. Our goal is to estimate the correctness of observations from a large crowd of unvetted sources with unknown reliability and various degrees of confidence rather than predict users' ratings or preferences of an item. Moreover, recommendation systems commonly assume a reasonable amount of good data is available to train their models while little is known about the data quality and the source reliability a priori in social sensing applications.

III. Confidence-aware Truth Estimation Problem in Social Sensing

In this section, we formulate the confidence-aware truth estimation problem in social sensing as an optimization problem (in the sense of maximum likelihood estimation). In particular, we consider a group of M sources, namely, S_1, S_2,...,S_M, who collectively report a set of N observations about the physical world, namely, C_1, C_2,..,C_N. Since we normally do not know the correctness of such observations a priori, we refer to them as *claims*. In this paper, we focus on the case of binary claims because the states of the physical environment in many applications can be represented by a set of statements that are either true or false. For example, in an application where the goal is to find potholes on city streets, each possible location may be associated with one claim that is true if a pothole presents at that location and false otherwise. Similarly, in an application that reports the free parking lot on campus, each parking lot may be associated with a claim that is true if that parking lot is free and false otherwise. In general, any statement about the physical world, such as "The bridge fell down","The building X is on fire", or "The suspect was captured" can be seen as a claim that is true if the statement is correct, and false if it is not. We assume, without loss of generality, that the "normal" state of each claim is negative (e.g., no potholes on streets and no free parking spots). Hence, sources report only when a positive value is encountered. Let S_i represent the i^{th} source and C_j represent the j^{th} claim. $C_j = 1$ if it is true and $C_j = 0$ if it is false. We define a *Sensing Matrix* SC to represent the relations between sources and claims, i.e., $S_iC_j = 1$ indicates that S_i reports C_j to be true, and $S_iC_j = 0$ otherwise.

In social sensing applications, human sources often express certain degree of confidence in their reported claims. To capture such confidence information, we define a *Confidence Matrix* W, where the element w_{ij} represents the degree of confidence S_i expresses on the claim C_j. Considering the difficulty of measuring the exact degree of confidence from human generated claims (e.g., text, images, etc.), we define w_{ij} as a discrete variable with K different values. In particular, $w_{ij} = k$ denotes that S_i reports the claim C_j to be true with a confidence degree of k, where $k = 1, ..., K$. In this paper, we explicitly consider the *source confidence information* and formulate a confidence-aware truth estimation problem in social sensing as follows.

First, let us define a few important terms that will be used in the problem formulation. We denote the reliability of source i as t_i, which is the probability a claim is true given the source S_i reports it. Formally t_i is given by:

$$t_i = P(C_j = 1|S_iC_j = 1) \tag{1}$$

Considering a source may make a claim with different degrees of confidence, we define t_i^k as the reliability of S_i when it reports a claim with a confidence degree of k, where $k = 1, ..., K$. Formally, t_i^k is given by:

$$t_i^k = P(C_j = 1|S_iC_j = 1, w_{ij} = k) \tag{2}$$

Therefore,

$$t_i = \sum_{k=1}^{K} t_i^k \times \frac{s_i^k}{s_i} \quad k = 1, ..., K \tag{3}$$

where s_i^k is the probability that S_i reports C_j with a confidence degree of k. For each source, s_i^k can be estimated using the Confidence Matrix W. Besides, considering different sources may make different number of claims, we denote the probability that S_i makes a claim by s_i. Formally, $s_i = P(S_iC_j = 1)$. Note that $s_i = \Sigma_{k=1}^{K} s_i^k$.

Let us further define $T_{i,k}$ to be the probability that S_i reports C_j to be true with a confidence degree of k, given that the claim is indeed true. Similarly, let $F_{i,k}$ denote the probability that S_i reports C_j to be true with a confidence degree of k, given that the claim is false. Formally, $T_{i,k}$ and $F_{i,k}$ are defined as follows:

$$\begin{aligned} T_{i,k} &= P(S_iC_j = 1, w_{ij} = k|C_j = 1) \\ F_{i,k} &= P(S_iC_j = 1, w_{ij} = k|C_j = 0) \end{aligned} \tag{4}$$

Using the Bayesian theorem, we can establish the relation between $T_{i,k}$, $F_{i,k}$ and t_i^k, s_i^k as follows:

$$T_{i,k} = \frac{t_i^k \times s_i^k}{d} \quad F_{i,k} = \frac{(1 - t_i^k) \times s_i^k}{1 - d} \tag{5}$$

where d represents the prior probability that a randomly chosen claim is true, which can be jointly estimated in our solution presented in next section.

Therefore, the confidence-aware truth estimation problem studied in this paper is presented as a maximum likelihood estimation (MLE) problem: given only the Sensing Matrix SC and Confidence Matrix W, our goal is to estimate likelihood of the correctness of each claim in C. Formally, we compute:

$$\forall j, 1 \le j \le N : P(C_j = 1|SC, W) \tag{6}$$

IV. A Confidence-aware Maximum Likelihood Estimation Approach

In this section, we solve the maximum likelihood estimation problem formulated in the previous section using the Expectation-Maximization (EM) algorithm. We start with a brief review of the EM algorithm and the mathematical formulation of our problem. We then derive the E and M steps of the proposed Confidence-Aware EM scheme and summarize it using the pseudocode.

A. Background and Mathematical Formulation

Intuitively, what the EM algorithm generally does is to iteratively estimate the values of the unknown parameters of a model and the values of the latent variables, which are not directly observable from the data. Such iterative process continues until the estimation results converge. To apply the EM algorithm to solve a MLE problem, we need to define a likelihood function $L(\theta; X, Z) = p(X, Z|\theta)$, where θ is the estimation parameter of the model, X is the observed data and Z is the set of latent variables. The EM algorithm finds the maximum likelihood estimate of θ and values of Z by iteratively performing two key steps: Expectation step (E-step) and Maximization step (M-step) that are given by:

$$\text{E-step: } Q(\theta|\theta^{(n)}) = E_{Z|x,\theta^{(n)}}[logL(\theta; x, Z)] \tag{7}$$

$$\text{M-step: } \theta^{(n+1)} = \arg\max_{\theta} Q(\theta|\theta^{(n)}) \tag{8}$$

For the MLE problem we formulated in Section III, the observed data is Sensing Matrix SC and the Confidence Matrix W. The estimation parameter is $\theta = (T_{1,k}, T_{2,k}, ..., T_{M,k}; F_{1,k}, F_{2,k}, ..., F_{M,k}; d)$ and $k = 1, 2, ..K$. $T_{i,k}$ and $F_{i,k}$ are defined in Equation (4) and d represents the prior probability of a randomly chosen claim to be true. We also need to define a vector of latent variables Z to indicate whether a claim is true or false. More specifically, we have a corresponding variable z_j for the j^{th} claim C_j such that: $z_j = 1$ if C_j is true and $z_j = 0$ otherwise. Most importantly, in order to incorporate different degrees of confidence a source may express on her/his claims into the MLE problem, we define a set of binary variables w_{ij}^k such that $w_{ij}^k = 1$ if $w_{ij} = k$ in Confidence Matrix W and $w_{ij}^k = 0$ otherwise. Therefore, the likelihood function of the confidence-aware truth estimation problem is given as:

$$
\begin{aligned}
& L(\theta; X, Z) = p(X, Z|\theta) \\
& = \prod_{j=1}^{N} \Big\{ \prod_{i=1}^{M} (T_{i,1})^{S_iC_j \ \&\& \ w_{ij}^1} \times ... \times (T_{i,K})^{S_iC_j \ \&\& \ w_{ij}^K} \\
& \times ((1 - \sum_{k=1}^{K} T_{i,k})^{1-S_iC_j}) \times d \times z_j \\
& + \prod_{i=1}^{M} (F_{i,1})^{S_iC_j \ \&\& \ w_{ij}^1} \times ... \times (F_{i,K})^{S_iC_j \ \&\& \ w_{ij}^K} \\
& \times ((1 - \sum_{k=1}^{K} F_{i,k})^{1-S_iC_j}) \times (1-d) \times (1-z_j) \Big\}
\end{aligned} \tag{9}
$$

where $S_iC_j = 1$ when source S_i reports C_j to be true and 0 otherwise and "&&" represents the "AND" logic for binary variables. The likelihood function represents the likelihood of the observed data (i.e., SC and W) and values of hidden variables (i.e., Z) given the estimation parameters (i.e., θ).

B. Deriving the E and M Steps

Given the above mathematical formulation, we derive E-step and M-step of the proposed Confidence-Aware EM scheme. First, we plug the likelihood function, given by Equation (9), into Equation (7) to derive the E-step as:

$$
\begin{aligned}
& Q(\theta|\theta^{(n)}) = E_{Z|X,\theta^{(n)}}[logL(\theta; X, Z)] \\
& = \sum_{j=1}^{N} \Big\{ p(z_j = 1|X_j, \theta^{(n)}) \times \sum_{i=1}^{M} \{ \sum_{k=1}^{K} (S_iC_j \ \&\& \ w_{ij}^k) \times logT_{i,k} \\
& + (1 - S_iC_j) log(1 - \sum_{k=1}^{K} T_{i,k}) + logd \} \\
& + p(z_j = 0|X_j, \theta^{(n)}) \times \sum_{i=1}^{M} \{ \sum_{k=1}^{K} (S_iC_j \ \&\& \ w_{ij}^k) \times logF_{i,k} \\
& + (1 - S_iC_j) log(1 - \sum_{k=1}^{K} F_{i,k}) + log(1-d) \Big\}
\end{aligned} \tag{10}
$$

We then define $Z(n, j) = p(z_j = 1|X_j, \theta^{(n)})$. It is the conditional probability of claim C_j to be true given the observed data X_j and current estimate of θ, where X_j represents the j^{th} column of both the Sensing Matrix SC and Confidence Matrix W. $Z(n, j)$ can be further expressed as:

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

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

$$
\begin{aligned}
& A(n, j) = p(X_j, \theta^{(n)}|z_j = 1) \\
& = \prod_{i=1}^{M} \prod_{k=1}^{K} (T_{i,k}^{(n)})^{S_iC_j \ \&\& \ w_{ij}^k} \times (1 - \sum_{k=1}^{K} (T_{i,k}^{(n)})^{1-S_iC_j}) \\
& B(n, j) = p(X_j, \theta^{(n)}|z_j = 0) \\
& = \prod_{i=1}^{M} \prod_{k=1}^{K} (F_{i,k}^{(n)})^{S_iC_j \ \&\& \ w_{ij}^k} \times (1 - \sum_{k=1}^{K} (F_{i,k}^{(n)})^{1-S_iC_j})
\end{aligned} \tag{12}
$$

$A(n, j)$ and $B(n, j)$ represent the conditional probability regarding observations of the claim C_j and current estimation of parameter θ, given that claim C_j is true and false respectively.

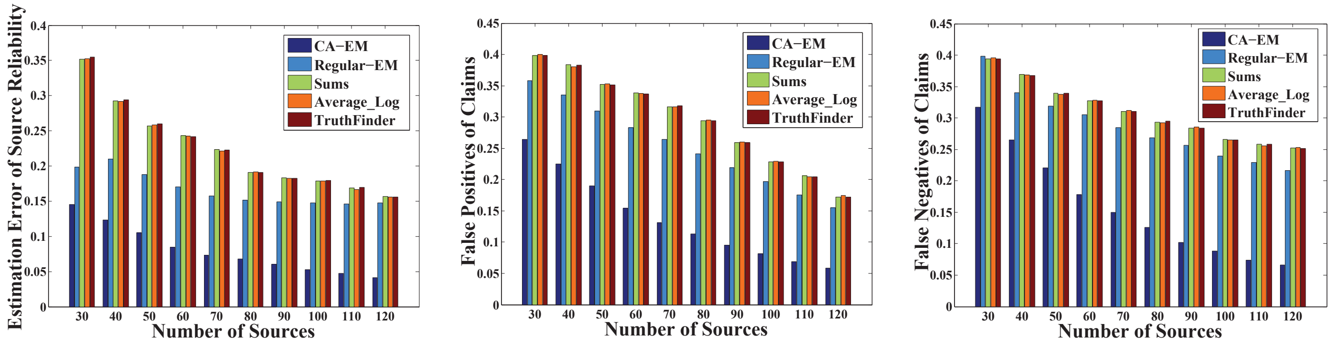
For the M-step, we set partial derivatives of $Q(\theta|\theta^{(n)})$ given by Equation (10) with respect to θ to 0 in order to get the optimal θ^* that maximizes Q function. In particular, solving $\frac{\partial Q}{\partial T_{i,k}} = 0$, $\frac{\partial Q}{\partial F_{i,k}} = 0$ and $\frac{\partial Q}{\partial d} = 0$, we can get expressions of the optimal $T_{i,k}^*$, $F_{i,k}^*$ and d^*:

$$
\begin{aligned}
T_{i,k}^{(n+1)} &= T_{i,k}^* = \frac{\Sigma_{j \in SW_i^k} Z(n, j)}{\Sigma_{j=1}^{N} Z(n, j)} \\
F_{i,k}^{(n+1)} &= F_{i,k}^* = \frac{\Sigma_{j \in SW_i^k} (1 - Z(n, j))}{N - \Sigma_{j=1}^{N} Z(n, j)} \\
d^{(n+1)} &= d^* = \frac{\Sigma_{j=1}^{N} Z(n, j)}{N}
\end{aligned} \tag{13}
$$

where N is the total number of claims in the Sensing Matrix SC and SW_i^k denotes the set of claims that source S_i reports with the confidence degree of k.

C. The Confidence-Aware EM Algorithm

In summary of the Confidence-Aware EM Algorithm derived in this section, the input is the Sensing Matrix SC and Confidence Matrix W obtained from social sensing data. The output is the maximum likelihood estimation of estimation parameters and values of latent variables, which can be used to compute both source reliability and the correctness of claims. The E-step and M-step of the Confidence-Aware EM algorithm are shown in Equation (11) and Equation (13) respectively. The convergence analysis has been done for EM scheme and it is beyond the scope of this paper [38]. We summarize the Confidence-Aware EM scheme in Algorithm 1.



(a) Source Reliability Estimation Accuracy (b) Claims Estimation: False Positives (c) Claims Estimation: False Negatives

Figure 1. Estimation Accuracy versus Number of Sources

Algorithm 1 Confidence-Aware EM Algorithm

Initialize θ ($T_{i,k} = s_i^k, F_{i,k} = 0.5 \times s_i^k$, d =Random number in $(0, 1)$
while $\theta^{(n)}$ does not converge **do**
  for $j = 1 : N$ **do**
    compute $Z(n, j)$ based on Equation (11)
  end for
  $\theta^{(n+1)} = \theta^{(n)}$
  for $i = 1 : M$ **do**
    compute $T_{i,k}^{(n+1)}, F_{i,k}^{(n+1)}, d^{(n+1)}$ based on Equation (13)
    update $T_{i,k}^{(n)}, F_{i,k}^{(n)}, d^{(n)}$ with $T_{i,k}^{(n+1)}, F_{i,k}^{(n+1)}, d^{(n+1)}$ in $\theta^{(n+1)}$
  end for
  $n = n + 1$
end while
Let Z_j^c = converged value of $Z(n, j)$
Let $T_{i,k}^c$ = converged value of $T_{i,k}^{(n)}$; $F_{i,k}^c$ = converged value of $F_{i,k}^{(n)}$; d^c = converged value of $d^{(n)}$
for $j = 1 : N$ **do**
  if $Z_j^c \geq 0.5$ **then**
    claim C_j is true
  else
    claim C_j is false
  end if
end for
for $i = 1 : M$ **do**
  calculate t_i^{k*} from $T_{i,k}^c$, $F_{i,k}^c$ and d^c based on Equation (4)
  calculate $t_i{}^*$ form t_i^{k*} based on Equation (3)
end for
Return the MLE on source reliability $t_i{}^*$ and corresponding judgment on the correctness of claim C_j.

V. Evaluation

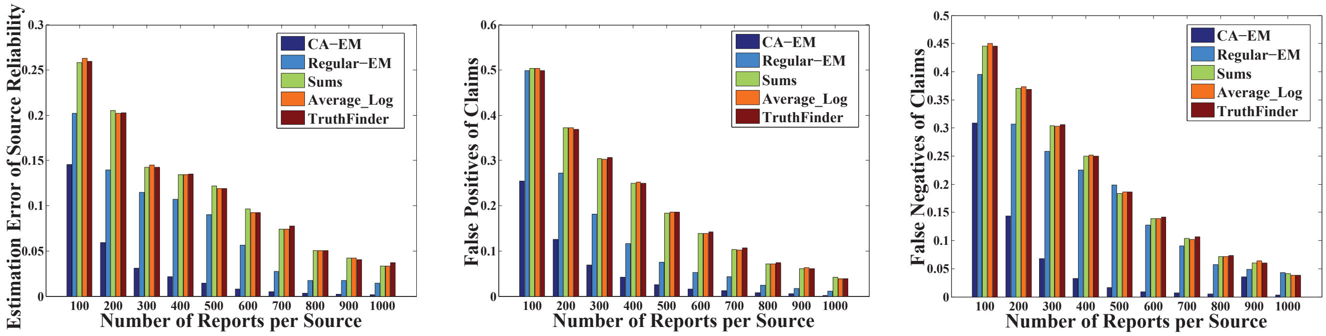
In this section, we carry out experiments to evaluate the performance of the proposed Confidence-Aware EM scheme (i.e., CA-EM) through both a simulation study and real world case studies in social sensing. We compare the CA-EM scheme with the-state-of-the-art baselines from the literature and showed that significant performance improvements can be achieved by the CA-EM scheme compared to the baselines.

A. Simulation Study

We begin our evaluation with an extensive simulation study of the proposed CA-EM scheme over different problem dimensions. We implement a simulator in Python 2.7 and compare the performance of the CA-EM scheme with the Regular EM in IPSN 12 [36] and other three state-of-the-art truth estimation schemes: Sums [15], Average_Log [21], and TruthFinder [42]. The simulator generates a random number of sources and claims. A random probability t_i is assigned to each source S_i representing her/his reliability (i.e., the ground truth probability S_i reports claims correctly). For source S_i, R_i reports are generated. Importantly, source S_i may report claim C_j with a certain degree of confidence k, $k = 1, ..., K$. We set $K = 3$ for the first three experiments and vary the value of K in the last experiment. For $K = 3$, we let t_i^H, t_i^M and t_i^L denote the source reliability with a confidence degree of *high*, *medium* and *low*. They are set to be uniformly distributed between (0.8,1), (0.5,0.8) and (0,0.5) respectively. In the evaluation, we compared the CA-EM with all baselines in terms of the estimation error of source reliability as well as the false positives and false negatives of the claim classification. The reported results are averaged over 100 experiments for all compared schemes.

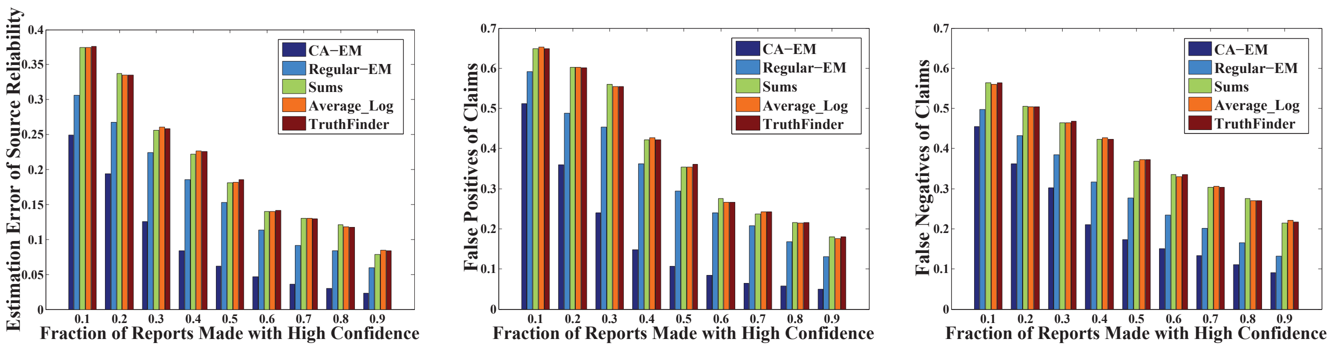
In the first experiment, we compare all schemes by varying the number of sources in the system. The number of reported claims was fixed at 2000, of which 1000 claims were true and 1000 were false. Since the non-EM schemes (i.e., Sums, Average_Log and TruthFinder) need the background prior d to be known as their input, we give the correct value of d (i.e., 0.5) to these schemes. The average number of reports made per source was set to 100. The fraction of reports made by a source with high, medium and low confidence was set to be equal (i.e., 1/3 for each). The number of sources was varied from 30 to 120. Results are shown in Figure 1. We observe that the CA-EM scheme outperforms all baselines by reducing both the estimation error on source reliability and false positives and false negatives on claim classification. We also note that the performance gain of the CA-EM scheme is significant as the number of sources changes in the system.

The second experiment compares the CA-EM scheme with all baselines when the average number of reports per source changes. We set the number of sources to 30. The average number of reports per source is varied from 100 to 1000. Other configurations are kept the same as the previous experiment. The results are shown in Figure 2. Observe that CA-EM outperforms all baselines in terms of both source reliability estimation accuracy and the claim classification accuracy as



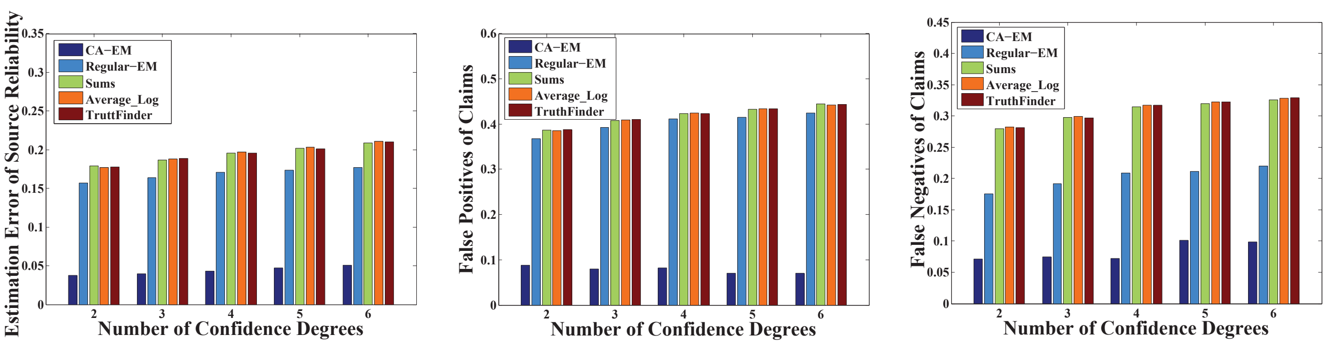
(a) Source Reliability Estimation Accuracy (b) Claims Estimation: False Positives (c) Claims Estimation: False Negatives

Figure 2. Estimation Accuracy versus Average Number of Reports per Source



(a) Source Reliability Estimation Accuracy (b) Claims Estimation: False Positives (c) Claims Estimation: False Negatives

Figure 3. Estimation Accuracy versus Fraction of Reports Made with High Confidence



(a) Source Reliability Estimation Accuracy (b) Claims Estimation: False Positives (c) Claims Estimation: False Negatives

Figure 4. Estimation Accuracy versus Number of Confidence Degrees

the average number of reports per source changes.

In the third experiment, we evaluate the performance of the CA-EM scheme and other schemes when the fraction of reports made with high confidence varies. We vary the fraction of high confidence reports from 0.1 to 0.9. The fractions of reports made with medium and low confidence were kept equal. The number of source is set to 30 and the average number of reports per source is set to 150. Reported results are shown in Figure 3. Observe that the CA-EM scheme achieves the best performance as compared to all baselines when the fraction of high confidence reports changes in the system.

In the last experiment, we evaluate the performance of all schemes by varying the number of confidence degrees (i.e., K) in the system. We vary the value of K from 2 to 6. The fraction of reports made with different degree of confidence are kept the same. The mean value of source reliability is constant over different K values. The average number of reports made per source is set to 300. Other configurations are the same as before. Reported results are shown in Figure 4. We observe that the CA-EM scheme consistently outperforms all baselines under different values of K. The performance gain of CA-EM compared to other baselines in the above experiments are achieved by judiciously exploring *different degrees of confidence* a source expresses on her/his claims.

This concludes our general simulations. In the next subsection, we will further evaluate the performance of CA-EM through several real-world case studies in social sensing.

B. Real Word Case Studies

In this subsection, we evaluate the CA-EM scheme using three real world case studies based on Twitter. Given Twitter is designed as an open data-sharing platform for average people, it creates an ideal scenario for unreliable content from unvetted human sources with various degrees of confidence. In our evaluation, we compare *CA-EM* to three representative baselines from current literature. The first baseline is *Voting*, which computes the data credibility simply by counting the number of times the same tweet is repeated on Twitter. The second baseline is the *Sums*, which also considers the differences in source reliability when it computes the data credibility scores [15]. The third baseline is the *Regular EM*, which was shown to outperform four current truth estimation schemes in social sensing [36].

In order to evaluate these schemes through real world case studies, we implemented them inside Apollo. Apollo was a fact-finding tool the authors have developed to capture tweets from many events of interest such as terrorist incidents, hurricanes, riots, civil unrest, and other natural and man-made disasters [3]. In particular, Apollo has: (i) a data collection component that allows users to collect tweets by specifying a set of keywords and/or geo-locations as filtering conditions and log the collected tweets; (ii) a data pre-processing component that clusters similar tweets into the same cluster by using micro-blog data clustering methods [29]. Using the meta-data output by the data pre-processing component, we generated the Sensing Matrix SC by taking the Twitter users as the data sources and the clusters of tweets as the the statements of user's observations, hence representing the *claims* in our model. The next step is to generate the Confidence Matrix W. For simplicity, we focus on the binary case here (i.e., $K = 2$). In particular, we use the following simple heuristics to roughly estimate the degree of confidence a user may express on a tweet: (i) if the tweet is an original tweet (i.e., not a retweet), it is of high confidence. Otherwise it is of low confidence. The hypothesis for this heuristic is that the first-hand information is often of higher confidence than the second-hand one (e.g., retweet). (ii) If the tweet contains a valid URL to an external source as the supporting evidence, it is of high confidence. Otherwise it is of low confidence. The hypothesis is that including external evidence normally indicates stronger confidence of users. (iii) The combination of the above two: if the tweet is an original tweet and contains a valid supporting URL, it is of high confidence. Otherwise, it is of low confidence. We note that the above heuristics are only the first approximations to estimate the degree of confidence a source may express in a tweet. The goal of this section is to provide a proof-of-concept demonstration of using these heuristics to estimate the source's confidence and validating the CA-EM scheme in real world social sensing applications. In future, we will explore more comprehensive text analysis techniques (e.g., natural language processing) to improve the accuracy of the source confidence estimation from tweets.

For the purposes of evaluation, we select three real world Twitter data traces of different sizes. The first trace was collected by Apollo during Boston Marathon bombings that happened on April 15, 2013 and subsequent shootings and manhunt events. The second was collected during and shortly after hurricane Sandy (the second-costliest hurricane in United States history) from New York and New Jersey in October/November 2012. The third one was collected from Cairo, Egypt during the violent unrest events that led to the resignation of the former Egyptian government in February 2011. These traces are summarized in Table I.

Trace	Boston Bombing	Hurricane Sandy	Egypt Unrest
Start Date	4/15/2013	11/2/2012	2/2/2011
Time duration	7 days	14 days	18 days
Physical Location	Boston and its suburbs	New York & New Jersey	Cairo
# of tweets	123,402	12,931	93,208
# of users tweeted	101,209	7,583	13,836

Table I
DATA STATISTICS OF THREE TRACES

We fed each data trace to the Apollo tool and executed all the compared truth estimation schemes. The output of these schemes was manually graded in each case to determine the credibility of the claims. Due to man-power limitations, we manually graded only the 100 top ranked claims by each scheme using the following rubric:

- *True claims:* Claims that are statements of a physical or social event, which is generally observable by multiple independent observers and corroborated by credible sources external to Twitter (e.g., mainstream news media).
- *Unconfirmed claims:* Claims that do not sastify the requirement of true claims.

We note that the unconfirmed claims may include the false claims and some possibly true claims that cannot be independently verified by external sources. Hence, our evaluation provides *pessimistic* performance bounds on the estimates.

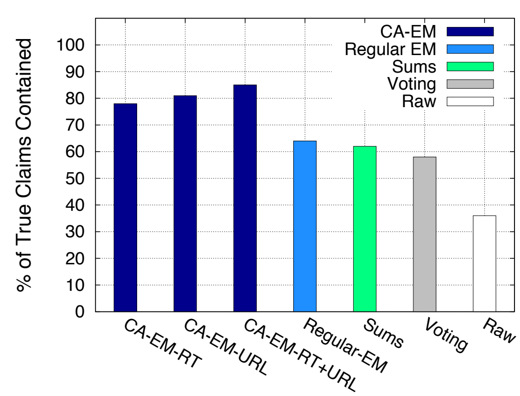


Figure 5. Evaluation on Boston Bombing Trace

Figure 5 shows the result for the Boston Marathon bombing trace. We observe that CA-EM schemes generally outperform the Regular EM scheme and other baselines in providing more true claims and suppressing the unconfirmed claims. This is achieved by explicitly incorporating different degree of confidence a source may express on the reported claims into the maximum likelihood estimation framework. The performance gains of CA-EM schemes compared to Regular EM are significant: *15%* to *22%* depending on the heuristics we used to generate the Confidence Matrix. It verifies the validity of using the CA-EM scheme to obtain more credible information in a real world social sensing application where sources are unvetted and likely to express various degrees of confidence about their claims. We also included the reference point called *Raw*, which indicates the average percentage of true claims in a random sample set of raw tweets.

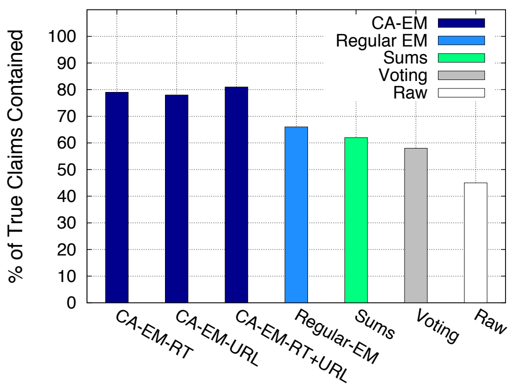


Figure 6. Evaluation on Hurricane Sandy Trace

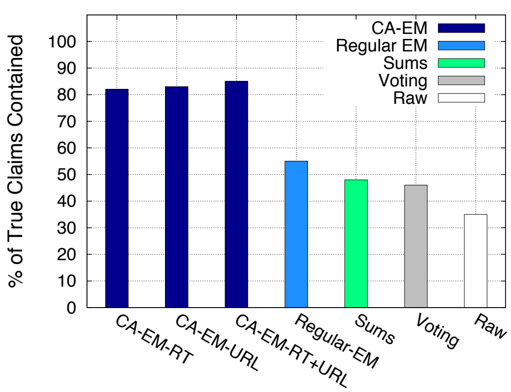


Figure 7. Evaluation on Egypt Unrest Trace

We repeated the above experiments on the Hurricane Sandy trace and Egypt Unrest trace. The results for Sandy trace are shown in Figure 6. In Figure 6, we consistently observe that the CA-EM schemes achieve the best performance compared to all baselines in increasing the number of true claims. Similar results are shown for the Egypt trace in Figure 7. The performance gain of the best performed CA-EM scheme in this case is even more significant (i.e., identifying 30% more true claims than the Regular EM).

VI. Discussions and Future Work

Sources are assumed to be independent in the current CA-EM scheme. However, dependency may exist between sources, especially when they are connected through the social networks. A recent effort has developed an effective model to address the source dependency in social sensing [34]. Also no correlations are assumed between claims in our framework. The claim correlation problem has been studied by the authors in a separate line of work [33]. Moreover, the aforementioned solutions on source dependency and claim correlation were developed under the same analytical framework as the CA-EM scheme. This allows the authors to quickly develop a more generalized confidence-aware truth estimation model that explicitly considers both the source dependency and claim correlation under a unified framework.

The current confidence estimation heuristics used in the Apollo system offers many opportunities for future improvement. The RT, URL, RT+URL are only first approximations. Authors plan to improve them by leveraging more comprehensive techniques (e.g., text mining, natural language processing, etc.) to estimate source's confidence from a deeper analysis of the tweet contents. Some recent efforts provide good insights into this direction by developing new methods to exploit the lexicon, syntax and semantics of data from Twitter [10], [30]. Moreover, the confidence estimation module is a plug-in of the proposed MLE framework, which gives us the flexibility to substitute it with a more refined one in the future.

The time dimension of the problem deserves more investigation. When the confidence of a source changes with large dynamics over time, how to best account for it in the MLE framework? A time-sensitive model is needed to better handle such dynamics. Recent work in fact-finding literature starts to develop a new category of streaming EM algorithms that quickly update the parameters of the maximum likelihood estimation using a recursive estimation approach [32]. Inspired by these results, the authors plan to develop similar real-time features of our CA-EM scheme to better capture the dynamics in the source's confidence. One key challenge is to design a nice tradeoff between estimation accuracy and computation complexity of the streaming algorithm. The authors are actively working on accommodating the above extensions.

VII. Conclusion

This paper presents a confidence-ware maximum likelihood estimation framework for explicitly considering *source confidence* to improve the truth estimation accuracy in social sensing applications. The approach can jointly estimate both the reliability of data sources and the credibility of their claims while exploiting the diversity of source confidence in their claims. The optimal solution is obtained by solving an expectation maximization problem where degrees of source confidence are encoded as link weights in a confidence matrix. We evaluated the proposed CA-EM scheme through both a simulation study and three real world case studies in social sensing. Evaluation results showed significant estimation accuracy improvements can be achieved by the CA-EM scheme compared to state-of-the-art baselines in the current literature.

REFERENCES

[1] T. Abdelzaher and D. Wang. Analytic challenges in social sensing. In *The Art of Wireless Sensor Networks*, pages 609–638. Springer, 2014.

[2] C. C. Aggarwal and T. Abdelzaher. Social sensing. In *Managing and Mining Sensor Data*, pages 237–297. Springer, 2013.

[3] Apollo-Toward Fact-finding for Social Sensing. http://apollo.cse.nd.edu/.

[4] I. Boutsis and V. Kalogeraki. Privacy preservation for participatory sensing data. In *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, volume 18, page 22, 2013.

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

[6] C. T. Chou, A. Ignjatovic, and W. Hu. Efficient computation of robust average of compressive sensing data in wireless sensor networks in the presence of sensor faults. *Parallel and Distributed Systems, IEEE Transactions on*, 24(8):1525–1534, 2013.

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

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

[9] M. Gupta and J. Han. Heterogeneous network-based trust analysis: a survey. *ACM SIGKDD Explorations Newsletter*, 13(1):54–71, 2011.

[10] M. Gupta, P. Zhao, and J. Han. Evaluating event credibility on twitter. In *SDM*, pages 153–164. SIAM, 2012.

[11] 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.

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

[13] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender systems: an introduction*. Cambridge University Press, 2010.

[14] A. Jøsang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decision support systems*, 43(2):618–644, 2007.

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

[16] N. D. Lane, S. B. Eisenman, M. Musolesi, E. Miluzzo, and A. T. Campbell. Urban sensing systems: opportunistic or participatory? In *Proceedings of the 9th workshop on Mobile computing systems and applications*, HotMobile '08, pages 11–16, New York, NY, USA, 2008. ACM.

[17] Z. Li, H. Shen, and K. Sapra. Leveraging social networks to combat collusion in reputation systems for peer-to-peer networks. *Computers, IEEE Transactions on*, 62(9):1745–1759, 2013.

[18] E. Miluzzo, N. D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. B. Eisenman, X. Zheng, and A. T. Campbell. Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*, pages 337–350. ACM, 2008.

[19] S. Nath. 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)*, 2012.

[20] L. M. Oliveira and J. J. Rodrigues. Wireless sensor networks: a survey on environmental monitoring. *Journal of communications*, 6(2):143–151, 2011.

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

[22] J. Pasternack and D. Roth. Generalized fact-finding (poster paper). In *World Wide Web Conference (WWW'11)*, 2011.

[23] J. Pasternack and D. Roth. Latent credibility analysis. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1009–1020. International World Wide Web Conferences Steering Committee, 2013.

[24] D. Pomerantz and G. Dudek. Context dependent movie recommendations using a hierarchical bayesian model. In *Advances in Artificial Intelligence*, pages 98–109. Springer, 2009.

[25] K. K. Rachuri, C. Mascolo, M. Musolesi, and P. J. Rentfrow. Sociablesense: exploring the trade-offs of adaptive sampling and computation offloading for social sensing. In *Proceedings of the 17th annual international conference on Mobile computing and networking*, MobiCom '11, pages 73–84, New York, NY, USA, 2011. ACM.

[26] 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.

[27] X. Sheng and Y.-H. Hu. Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks. *Signal Processing, IEEE Transactions on*, 53(1):44–53, 2005.

[28] M. Srivastava, T. Abdelzaher, and B. K. Szymanski. Human-centric sensing. *Philosophical Transactions of the Royal Society*, 370(1958):176–197, January 2012.

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

[30] D. Tang, F. Wei, B. Qin, T. Liu, and M. Zhou. Coooolll: A deep learning system for twitter sentiment classification. *SemEval 2014*, page 208, 2014.

[31] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456. ACM, 2011.

[32] D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal. Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications. In *The 33rd International Conference on Distributed Computing Systems (ICDCS'13)*, July 2013.

[33] D. Wang, T. Abdelzaher, L. Kaplan, R. Ganti, S. Hu, and H. Liu. Exploitation of physical constraints for reliable social sensing. In *The IEEE 34th Real-Time Systems Symposium (RTSS'13)*, 2013.

[34] D. Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. C. Aggarwal, R. Ganti, et al. Using humans as sensors: an estimation-theoretic perspective. In *Proceedings of the 13th international symposium on Information processing in sensor networks*, pages 35–46. IEEE Press, 2014.

[35] D. Wang, L. Kaplan, T. Abdelzaher, and C. C. Aggarwal. 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)*, June 2012.

[36] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher. On truth discovery in social sensing: A maximum likelihood estimation approach. In *The 11th ACM/IEEE Conference on Information Processing in Sensor Networks (IPSN 12)*, April 2012.

[37] X. Wang, M. Fu, and H. Zhang. Target tracking in wireless sensor networks based on the combination of kf and mle using distance measurements. *Mobile Computing, IEEE Transactions on*, 11(4):567–576, 2012.

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

[39] Y.-C. Wu, Q. Chaudhari, and E. Serpedin. Clock synchronization of wireless sensor networks. *Signal Processing Magazine, IEEE*, 28(1):124–138, 2011.

[40] L. Xiao, S. Boyd, and S. Lall. A scheme for robust distributed sensor fusion based on average consensus. In *Information Processing in Sensor Networks, 2005. IPSN 2005. Fourth International Symposium on*, pages 63–70. IEEE, 2005.

[41] 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.

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

[43] D. Yu, H. Huang, T. Cassidy, H. Ji, C. Wang, S. Zhi, J. Han, C. Voss, and M. . Magdon-Ismail. The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding. In *The 25th International Conference on Computational Linguistics*, 2014.

[44] S. Zhao, L. Zhong, J. Wickramasuriya, and V. Vasudevan. Human as real-time sensors of social and physical events: A case study of twitter and sports games. *CoRR*, abs/1106.4300, 2011.

[45] Z. Zhao, J. Cheng, and W. Ng. Truth discovery in data streams: A single-pass probabilistic approach. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1589–1598. ACM, 2014.