2015 IEEE 12th International Conference on Mobile Ad Hoc and Sensor Systems

Towards Time-Sensitive Truth Discovery in Social Sensing Applications

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

***Abstract*—This paper develops a new principled framework for exploiting time-sensitive information to improve the truth discovery accuracy in social sensing applications. This work is motivated by the emergence of social sensing as a new paradigm of collecting observations about the physical environment from humans or devices on their behalf. These observations may be true or false, and hence are viewed as binary claims. A fundamental problem in social sensing applications lies in ascertaining the correctness of claims and the reliability of data sources. We refer to this problem as *truth discovery*. Time is a critical dimension that needs to be carefully exploited in the truth discovery solutions. In this paper, we develop a new time-sensitive truth discovery scheme that explicitly incorporates the *source responsiveness* and the *claim lifespan* into a rigorous analytical framework. The new truth discovery scheme solves a maximum likelihood estimation problem to determine both the claim correctness and the source reliability. We compare our time-sensitive scheme with the state-of-the-art baselines through an extensive simulation study and a real world case study. The evaluation results showed that our new scheme outperforms all compared baselines and significantly improves the truth discovery accuracy in social sensing applications.**

***Keywords*-Social Sensing, Time-sensitive, Truth Discovery, Source Responsiveness, Claim Lifespan, Maximum Likelihood Estimation, Expectation Maximization**

I. Introduction

This paper investigates the exploitation of time-sensitive information to improve the truth discovery accuracy in social sensing applications. Social sensing has emerged as a new paradigm of collecting observations about the physical environment from humans or devices on their behalf [21]. This paradigm is motivated by the proliferation of various sensors in the possession of common individuals and the popularity of social networks that enable massive information dissemination opportunities. For example, drivers may contribute data through their smartphones to report the state of traffic congestion at various locales [9]. Alternatively, survivors may contribute data to online social media (e.g., Twitter, Facebook, Flickr) to document the damage in the aftermath of a natural disaster [24]. These observations may be true or false, and hence are viewed as binary *claims*. A fundamental problem in social sensing applications lies in accurately ascertaining the correctness of claims and the reliability of data sources. We refer to this problem as *truth discovery*.

Previous works in sensor network [25]–[28], information fusion [12], [20] and data mining [32], [34] have made significant advances to address the truth discovery problem in social sensing. However, the *time* dimension of the problem has not been fully exploited in these works. In particular, two pieces of time-sensitive information on both sources and claims (i.e., *source responsiveness* and *claim lifespan*) were largely missing in the state-of-the-art solutions. Ignorance of such time-sensitive information can easily lead to sub-optimal solutions to the truth discovery problem. There are two major reasons. First, sources may report claims of the same event with different *degrees of responsiveness*. Thus, a source who provides fast and original claims should be treated differently from a source who generates delayed and repeated claims. However, current truth discovery solutions simply ignored such difference in source responsiveness [27], [32]. Second, the reported events may last for different amount of time and claims associated with these events may have different lifespans. A claim with a longer lifespan accumulates more observations and is often considered to be more credible than a claim with a shorter lifespan in current literature [22], [28]. However, this is not always true. For example, a local emergency report lasting for a few hours can be more credible than a popular rumor lasting for a few days. As we observe in our experiments, if we totally ignored differences in claim lifespan, correct claims with short lifespans were easily mis-classified as false negatives in the truth discovery results.

Important challenges exist when we exploit the time-sensitive information to improve the truth discovery accuracy in social sensing. First, social sensing is designed as an open data collection paradigm where the reliability of sources and the correctness of claims are often *unknown* a priori. Normally, all we have are massive noisy observations from a large crowd of unvetted sources. Additionally, the source responsiveness and claim lifespan are closely related with the source reliability and claim correctness in non-trivial ways. However, how to develop a principled framework that can explicitly and rigorously incorporate such time-sensitive information into the truth discovery solutions remains a critical problem to be solved.

In this paper, we develop a time-sensitive truth discovery scheme that explicitly incorporates source responsiveness and claim lifespan into a maximum likelihood estimation framework. In particular, a Time-Sensitive Expectation Max-

978-1-4673-9101-6/15 $31.00 © 2015 IEEE
DOI 10.1109/MASS.2015.39

imization (TS-EM) algorithm is developed to assign true values to claims and reliability to sources more accurately by exploiting the time-sensitive information on both sources and claims. We evaluate our time-sensitive truth discovery scheme through an extensive simulation study and a real world case study in social sensing. The evaluation results show that our new scheme outperforms the previous works that ignore the time-sensitive information and other state-of-the-art baselines and significantly improves the truth discovery accuracy. The results of this paper are important because they allow social sensing applications to accurately estimate the correctness of claims and the reliability of sources by incorporating the time-sensitive information into a principled framework. To summarize, our contributions are as follows:

- To the best of our knowledge, we are the first to explicitly consider the time-sensitive information (i.e., source responsiveness and claim lifespan) in the truth discovery problem in social sensing.
- We develop a principled framework that allows us to derive an optimal solution (in the sense of maximum likelihood estimation) for the time-sensitive truth discovery problem.
- We show non-trivial performance gains achieved by our time-sensitive scheme (i.e., our scheme increased the claim classification precision by 17% compared to state-of-the-art baselines in a real world case study.).

The rest of this paper is organized as follows: we discuss the related work in Section II. In Section III, we present the new time-sensitive truth discovery model for social sensing applications. The proposed maximum likelihood estimation framework and the expectation maximization solution is presented in Section IV. Evaluation results are presented in Section V. Finally, we conclude the paper in Section VI.

II. Related Work

Social sensing has emerged as a new act of collecting sensory measurements about the physical world from human sources or devices on their behalf [21]. Some early applications include CenWits [11], CabSense [18], and BikeNet [7]. More recent applications in social sensing start to address challenges such as preserving privacy of participants [5], balancing the tradeoffs between sensing quality and costs [23], resolving information collision [29], and building general models in sparse and multi-dimensional social sensing spaces [3]. An emerging and critical question about data reliability arises when the data in social sensing applications are collected by humans whose "reliability" is not known [1]. Some truth discovery techniques have been developed to address this problem but they did not fully exploit the time dimension of the problem in their solutions [27], [28]. In this paper, we develop a time-sensitive truth discovery scheme that explicitly exploits the time-sensitive information (i.e., source responsiveness and claim lifespan) in social sensing and significantly improves the truth discovery accuracy.

In data mining and machine learning literature, there exists a good amount of work on the topics of *fact-finding* that jointly compute the source reliability and claim credibility. *Hubs and Authorities* [13] established a basic fact-finding model based on linear assumptions to compute scores for sources and claims they asserted. Yin et al. introduced *TruthFinder* as an unsupervised fact-finder for trust analysis on a providers-facts network [31]. Other fact-finders enhanced these basic frameworks by incorporating analysis on properties or dependencies within claims and sources [20]. More recently, new fact-finding algorithms have been designed to address the background knowledge [16], and multi-dimensional aspects of the problem [33]. In this paper, we use the insights from the above work and develop a new estimation scheme to solve the time-sensitive truth discovery problem in social sensing applications.

Maximum likelihood estimation (MLE) technique has been widely used in sensor network community to solve estimation and information fusion problems [14], [17], [30]. For example, Wang et al. proposed a MLE based target tracking approach to solve the instability problem and offer superior tracking performance in wireless sensor networks [30]. Pereira et al. presented a maximum likelihood estimation algorithm to solve a distributed parameter estimation problem in unreliable sensor networks [17]. Leng et al. built a maximum likelihood estimator to jointly estimate the clock offset, clock skew and fixed delay in sensor networks [14]. However, the estimation variables in the above work are mostly continuous and the sensors are physical sensors. In contrast, we focus on estimating a set of *binary variables* that represent either true or false statements from human sensors. The MLE problem we studied is actually more challenging due to the discrete nature of the estimated variables and the non-trivial complexity of modeling *humans as sensors* in social sensing.

Finally, our work is also related with reputation and trust systems that are designed to assess the reliability of sources (e.g., the quality of providers). eBay is a typical reputation system based on a homogeneous peer-to-peer network structure, which allows participants to rate each other after each pair of them conduct a transaction [10]. Alternatively, Amazon on-line review system represents another type of reputation system based on a heterogeneous network structure, where different sources offer reviews on products (or brands, companies) they experienced [8]. Recent work has also investigated the consistency of reports to estimate and revise trust scores in reputation systems [12]. However, in social sensing, we normally do not have enough history data to compute the converged reputation scores of sources due to the short-lived sensing campaigns. Instead, this paper presents a maximum likelihood estimation approach that jointly estimates both source reliability and the credibility of

data based on the observations collected from applications.

III. Time-Sensitive Truth Discovery Problem in Social Sensing

In this section, we formulate the time-sensitive truth discovery problem in social sensing as a maximum likelihood estimation problem. Consider a scenario where a group of M sources, namely, $S_1, S_2, ..., S_M$, who report a set of N observations about the physical environment, namely, $C_1, C_2, ..., C_N$. Since the correctness of these observations are normally not known to the application in advance, we refer to these observations as *claims*. In this paper, we focus on *binary* claims since the states of the physical world in many social sensing applications can be represented by a set of statements that are either true or false. For example, in an application that reports the offensive graffiti on campus walls, each location may be associated with a claim that is true if the offensive graffiti is present and false otherwise. In general, any statement about the physical environment, such as "A car accident happened on Main street", "Heavy gunfire broke out in Union Square" and "The building X is on fire" can be thought of as a binary claim that is true if the statement is correct, and false otherwise. We assume, without loss of generality, that the default state of each claim is negative (e.g., no graffiti on the wall). Hence, sources only report when the positive state of the claim 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 $C_j = 0$ otherwise. We define a *Sensing Matrix* SC, where $S_iC_j = 1$ when source S_i reports that claim C_j is true, and $S_iC_j = 0$ otherwise.

Furthermore, we need to incorporate the time-sensitive information (i.e., source responsiveness and claim lifespan) into our model. To capture the source responsiveness information, we define a *Responsiveness Matrix* R, where the element r_{ij} represents the degree of responsiveness of S_i when it reports the claim C_j. Specifically, r_{ij} is a discrete variable with K different values representing K possible degrees of source responsiveness (e.g., immediate, fast, slow). To capture the lifespan information, we define a *Lifespan Vector* W, where the element w_j represents the lifespan of claim C_j. Specifically, w_j is a discrete variable with L different values representing L categories of claim lifespan (e.g., short, medium, long).

We formulate the time-sensitive truth discovery 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 S_i by t_i, which is the probability that a claim is correct given that source S_i reported it. Formally, t_i is given by:

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

Considering a source S_i may report claims with different degrees of responsiveness and the reported claims may have different lifespans, we define $t_i^{k,l}$ as the reliability of S_i when it reports a claim with a responsiveness degree of k and the reported claim has a lifespan of l, where $k = 1, ..., K$, $l = 1, ..., L$. Formally, $t_i^{k,l}$ is given by:

$$t_i^{k,l} = \Pr(C_j = 1, w_j = l | S_iC_j = 1, r_{ij} = k) \tag{2}$$

Hence,

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

where $s_i^{k,l}$ is the probability that S_i reports C_j with a responsiveness degree of k and C_j has a lifespan of l, Formally, $s_i^{k,l} = \Pr(S_iC_j = 1, r_{ij} = k, w_j = l)$. Note that the probability that S_i reports a claim is: $s_i = \Sigma_{l=1}^{L} \Sigma_{k=1}^{K} s_i^{k,l}$.

Let us further define $T_i^{k,l}$ to be the (unknown) probability that S_i reports C_j (of lifespan l) with a responsiveness degree of k, given that the claim is indeed true. Similarly, let $F_i^{k,l}$ denote the (unknown) probability that S_i reports C_j (of lifespan l) with a responsiveness degree of k, given that the claim is false. Formally, $T_i^{k,l}$ and $F_i^{k,l}$ are defined as follows:

$$\begin{aligned} T_i^{k,l} &= \Pr(S_iC_j = 1, r_{ij} = k | C_j = 1, w_j = l) \\ F_i^{k,l} &= \Pr(S_iC_j = 1, r_{ij} = k | C_j = 0, w_j = l) \end{aligned} \tag{4}$$

Using the Bayes theorem, we can establish the relationship between $T_i^{k,l}$, $F_i^{k,l}$ and $t_i^{k,l}$, $s_i^{k,l}$ as follows:

$$T_i^{k,l} = t_i^{k,l} \times s_i^{k,l} / d^l, \quad F_i^{k,l} = (1 - t_i^{k,l}) \times s_i^{k,l} / (1 - d^l) \tag{5}$$

where d^l is the prior probability that a randomly chosen claim with a lifespan of l is true (i.e., $d^l = \Pr(C_j = 1, w_j = l)$). The introduced notations are summarized in Table I.

Table I
The Summary of Notations

Description	Notation
Set of Sources	S
Set of Claims	C
Sensing Matrix	SC
Responsiveness Matrix	R
Life Span Vector	W
Report Probability	$s_i^{k,l} = \Pr(S_iC_j = 1, r_{ij} = k, w_j = l)$
Source Reliability	$t_i^{k,l} = \Pr(C_j = 1, w_j = l \| S_iC_j = 1, r_{ij} = k)$
Correctness Probability	$T_i^{k,l} = \Pr(S_iC_j = 1, r_{ij} = k \| C_j = 1, w_j = l)$
Error Probability	$F_i^{k,l} = \Pr(S_iC_j = 1, r_{ij} = k \| C_j = 0, w_j = l)$

Therefore, the time-sensitive truth discovery problem studied in this paper can be formulated as a maximum likelihood estimation (MLE) problem: given the Sensing Matrix SC, Responsiveness Matrix R and Lifespan Vector W, we aim at estimating the likelihood of the correctness

of each claim and reliability of each source. Formally, we compute:

$$\forall j, 1 \leq j \leq N : \Pr(C_j = 1 | SC, R, W)$$
$$\forall i, 1 \leq i \leq M : \Pr(C_j = 1 | S_i C_j = 1) \quad (6)$$

IV. A Time-Sensitive Maximum Likelihood Estimation Approach

In this section, we solve the time-sensitive truth discovery problem formulated in Section III by developing a Time-Sensitive Expectation-Maximization (TS-EM) algorithm.

A. Building The Likelihood Function

EM is an optimization scheme that is commonly used to solve the MLE problem where unobserved latent variables exist in the model [6]. Specifically, it iterates between two key steps: expectation step (E-Step) and maximization step (M-step). In E-step, it computes the expectation of the log-likelihood function based on the current estimates of the model parameters. In M-step, it computes the new estimates of the model parameters that maximize the expected log-likelihood function in E-step. The two steps of EM are shown as follows:

$$\text{E-step: } Q(\theta|\theta^{(n)}) = E_{Z|x,\theta^{(n)}}[logL(\theta; x, Z)] \quad (7)$$
$$\text{M-step: } \theta^{(n+1)} = \arg\max_{\theta} Q(\theta|\theta^{(n)}) \quad (8)$$

where $L(\theta; X, Z) = \Pr(X, Z|\theta)$ is the likelihood function, θ is the estimation parameter of the model, X is the observed data and Z is a set of latent variables.

Now let us consider how to solve the MLE problem we formulated in the previous section by developing a time-sensitive EM scheme. First, we need to define the likelihood function of the MLE problem. In particular, the observed data X in our problem is the Sensing Matrix SC, the Responsiveness Matrix R and the Lifespan Vector W. The estimation parameter vector is defined as $\theta = (T_1^{k,l}, T_2^{k,l}, ..., T_M^{k,l}; F_1^{k,l}, F_2^{k,l}, ..., F_M^{k,l}; d^l)$ where $k = 1, ..., K,\ l = 1, ..., L$ and $T_i^{k,l}$, $F_i^{k,l}$ and d^l are defined in Equation (4) and (5). Furthermore, we need to define a vector of latent variables Z to indicate whether a claim is true or false. More specially, we have a corresponding variable z_j^l for claim C_j (whose lifespan category is l) such that $z_j^l = 1$ if C_j is true and $z_j^l = 0$ otherwise. Additionally, we define a set of binary indication variables r_{ij}^k such that $r_{ij}^k = 1$ if $r_{ij} = k$ in Responsiveness Matrix R and $r_{ij}^k = 0$ otherwise. Similarly, we define another set of binary indication variables w_j^l such that $w_j^l = 1$ if $w_j = l$ in Lifespan Vector W and $w_j^l = 0$ otherwise. Hence, the likelihood function of time-sensitive truth discovery problem can be written given as:

$$\begin{aligned}
&L(\theta; X, Z) = \Pr(X, Z|\theta) \\
&= \prod_{j=1}^{N} \Bigg\{ \prod_{l=1}^{L} \Bigg[\prod_{i=1}^{M} \prod_{k=1}^{K} (\alpha_{i,j}^{k,l})^{S_i C_j \ \&\& \ r_{ij}^k \ \&\& \ w_j^l} \\
&\times (\alpha_{i,j}^{k,l})^{(1-S_i C_j) \ \&\& \ w_j^l} \times d^l \times z_j^l \Bigg] \\
&+ \prod_{l=1}^{L} \Bigg[\prod_{i=1}^{M} \prod_{k=1}^{K} (\alpha_{i,j}^{k,l})^{S_i C_j \ \&\& \ r_{ij}^k \ \&\& \ w_j^l} \\
&\times (\alpha_{i,j}^{k,l})^{(1-S_i C_j) \ \&\& \ w_j^l} \times (1 - d^l) \times (1 - z_j^l) \Bigg] \Bigg\}
\end{aligned} \quad (9)$$

where $S_i C_j = 1$ when source S_i reports C_j to be true and 0 otherwise. The "&&" represents the "AND" logic for binary variables. The $\alpha_{i,j}^{k,l}$ are defined as follows:

$$\alpha_{i,j}^{k,l} = \begin{cases} T_i^{k,l} & if\ S_i C_j = 1, r_{i,j}^k = 1, w_j^l = 1, z_j^l = 1 \\ (1 - \sum_{k=1}^{K} T_i^{k,l}) & if\ S_i C_j = 0, w_j^l = 1, z_j^l = 1 \\ F_i^{k,l} & if\ S_i C_j = 1, r_{i,j}^k = 1, w_j^l = 1, z_j^l = 0 \\ (1 - \sum_{k=1}^{K} F_i^{k,l}) & if\ S_i C_j = 0, w_j^l = 1, z_j^l = 0 \end{cases} \quad (10)$$

The likelihood function represents the likelihood of the observed data (i.e., SC, R and W) and the values of hidden variables (i.e., Z) given the estimation parameters (i.e., θ).

B. Time-Sensitive Expectation Maximization

Given the above mathematical formulation, we derive E and M steps of the proposed TS-EM scheme. First, we derive the Q function for the E-step given by Equation (7) using the likelihood function derived in Equation (9). The E-step is given as follows:

$$
\begin{aligned}
&Q(\theta|\theta^{(n)}) = E_{Z|X,\theta^{(n)}}[logL(\theta; X, Z)] \\
&= \sum_{j=1}^{N} \Bigg\{ \sum_{l=1}^{L} \Pr(z_j^l = 1|X_j^l, \theta^{(n)}) \\
&\times \Bigg[\sum_{i=1}^{M} \sum_{k=1}^{K} (S_i C_j \ \&\& \ r_{ij}^k \ \&\& \ w_j^l) \times log\alpha_{i,j}^{k,l} \\
&+ \Big((1 - S_i C_j) \ \&\& \ w_j^l \Big) \times log\alpha_{i,j}^{k,l} + logd^l \Bigg] \\
&+ \sum_{l=1}^{L} \Pr(z_j^l = 0|X_j^l, \theta^{(n)}) \\
&\times \Bigg[\sum_{i=1}^{M} \sum_{k=1}^{K} (S_i C_j \ \&\& \ r_{ij}^k \ \&\& \ w_j^l) \times log\alpha_{i,j}^{k,l} \\
&+ \Big((1 - S_i C_j) \ \&\& \ w_j^l \Big) \times log\alpha_{i,j}^{k,l} + log(1 - d^l) \Bigg] \Bigg\}
\end{aligned} \tag{11}
$$

Note that in the Q function, the estimation parameters are represented by $\alpha_{i,j}^{k,l}$ which is defined in Equation (10). $\alpha_{i,j}^{k,l}$ represents different parameters under different conditions.

For the M-step, in order to get the optimal θ^* that maximizes Q function, we set partial derivatives of $Q(\theta|\theta^{(n)})$ given by Equation (11) with respect to θ to 0. In particular, we get the solutions of $\frac{\partial Q}{\partial T_i^{k,l}} = 0$, $\frac{\partial Q}{\partial F_i^{k,l}} = 0$ and $\frac{\partial Q}{\partial d^l} = 0$ in each iteration, we can get optimal estimations for the next iteration (i.e., $(T_i^{k,l})^{(n+1)}$, $(F_i^{k,l})^{(n+1)}$ and $(d^l)^{(n+1)}$):

$$
\begin{aligned}
(T_i^{k,l})^{(n+1)} &= \frac{\Sigma_{j \in SW_i^{k,l}} \Pr(z_j^l = 1|X_j^l, \theta^{(n)})}{\Sigma_{j \in C^l} \Pr(z_j^l = 1|X_j^l, \theta^{(n)})} \\
(F_i^{k,l})^{(n+1)} &= \frac{\Sigma_{j \in SW_i^{k,l}} (1 - \Pr(z_j^l = 1|X_j^l, \theta^{(n)}))}{\Sigma_{j \in C^l} (1 - \Pr(z_j^l = 1|X_j^l, \theta^{(n)}))} \\
(d^l)^{(n+1)} &= \frac{\Sigma_{j \in C^l} \Pr(z_j^l = 1|X_j^l, \theta^{(n)})}{|C^l|}
\end{aligned} \tag{12}
$$

where $SW_i^{k,l}$ is the set of claims (with lifespan l) that source S_i reports with the responsiveness degree of k. We also define C^l as the set of claims whose lifespan is l.

C. Summary of The Time-Sensitive EM Algorithm

In summary, the input of the TS-EM algorithm is the Sensing Matrix SC, Responsiveness Matrix R and Lifespan Vector W obtained from the social sensing data. The output is the maximum likelihood estimation of estimation parameters and latent variables. The estimation results can be used to compute both source reliability and claim correctness. We summarize the TS-EM scheme in Algorithm 1.

Algorithm 1 Time-Sensitive EM Algorithm

Input: Sensing Matrix SC, Responsiveness Matrix R, Lifespan Vector W
Output: Estimations of Source Reliability and Claim Correctness

```
Initialize θ (T_i^{k,l} = s_i^{k,l}, F_i^{k,l} = 0.5 × s_i^{k,l}, d^l =Random number in (0, 1))
n = 0
repeat
    n = n + 1
    for Each l ∈ {1, 2, ..., L} do
        for Each j ∈ C do
            compute Pr(z_j^l = 1|X_j^l, θ^(n))
        end for
        for Each i ∈ S do
            compute (T_i^{k,l})^(n), (F_i^{k,l})^(n), (d^l)^(n)
        end for
    end for
until θ^(n) and θ^(n−1) converge
Let (Z_j^l)^c = converged value of Pr(z_j^l = 1|X_j^l, θ^(n))
for Each l ∈ {1, 2, ..., L} do
    for Each j ∈ C do
        if (Z_j^l)^c ≥ 0.5 then
            claim C_j^l is true
        else
            claim C_j^l is false
        end if
    end for
    for Each i ∈ S do
        calculate (t_i^{k,l})* from converge values of (T_i^{k,l}), (F_i^{k,l}) and (d^l) based on Equation (5)
        calculate t_i* form (t_i^{k,l})* based on Equation (3)
    end for
end for
```

V. Evaluation

In this section, we evaluate the proposed Time-Sensitive EM scheme (TS-EM) through a simulation study and a real world case study. The experimental results show that the TS-EM outperforms the state-of-the-art baselines and significantly improves the truth discovery accuracy in social sensing applications.

A. Simulation Study

In this subsection we carry out an extensive simulation study to evaluate the performance of the TS-EM scheme over different problem dimensions. We focus on two performance metrics: (i) the accuracy of source reliability estimation and (ii) the accuracy of claim classification (i.e., false positives and false negatives). We compared the performance of TS-EM with the Regular-EM in IPSN 12 [28] and other three state-of-the-art baselines: Sums [13], Average_Log [15] and TruthFinder [32]. We built a social sensing simulator in Python 2.7. The simulator generates a random number of sources and claims. For source S_i, a random probability t_i is assigned to represent its reliability (i.e., the ground truth probability that S_i reports correct claims). Based on the reliability t_i, source S_i reports R_i claims. Importantly, source S_i reports a claim with certain degree of responsiveness k, $k = 1, ..., K$ and claim C_j has its unique category of lifespan l, $l = 1, .., L$. For the first two experiments, we set

$K = 3$ and $L = 3$ and evaluate the performance of TS-EM by changing different parameters of the model. For the last two experiments, we evaluate the TS-EM scheme by varying the value of K and L respectively. The reported results are the average of 100 experiments.

In the first experiment, we compare the TS-EM scheme with all baselines by varying the number of sources in our application. We set the total number of claims to 3000, of which 1500 claims were true and 1500 were false. The average number of reports per source was set to 300. K and L were both set to 3. The claims and reports are uniformly distributed between different categories of lifespan and degrees of source responsiveness respectively. In this experiment, we changed the number of sources from 30 to 120. Results are shown in Figure 1. We observe that TS-EM outperforms all baselines in terms of both source reliability estimation accuracy (i.e., smaller estimation errors) and the claim classification accuracy (i.e., smaller false positives/negatives). We also note that the performance improvement of TS-EM is significant as the number of sources in the system changes.

In the second experiment, we evaluate the performance of all schemes when changing the average number of reports made per source. In this experiment, the number of sources was fixed at 30. We varied the average number of reports per source from 150 to 1500. For other experiment parameters, we kept them the same as the first experiment. Reported results are shown in Figure 2. We observe that TS-EM scheme achieves the smallest estimation error on source reliability and the least false positives and false negatives on claim classification among all schemes we compared.

In the third experiment, we evaluate the performance of TS-EM and other baselines by varying the number of possible degrees of source responsiveness (i.e., K). We set the number of sources to 30 and the number of reports made per source to 300. The number of possible categories of claim lifespan was set to 3 (i.e., $L = 3$). We varied K from 2 to 6. For other parameters, we kept them the same as previous experiments. The results are shown in Figure 3. We observe that TS-EM scheme consistently outperforms all baselines under different values of K in terms of both source reliability estimation accuracy and claim classification accuracy.

In the last experiment, we evaluate the performance of all schemes by varying the number of possible categories of claim lifespan (i.e., L). In this experiment, we set the number of sources to 30 and the number of reports made per source to 300. The number of source responsiveness degree was set to 3 (i.e., $K = 3$). We varied L from 2 to 6. We kept other experiment configurations the same as before. Results are shown in Figure 4. We observe that the TS-EM scheme continues to outperform all baselines under different values of L in terms of both source reliability estimation accuracy and claim classification accuracy.

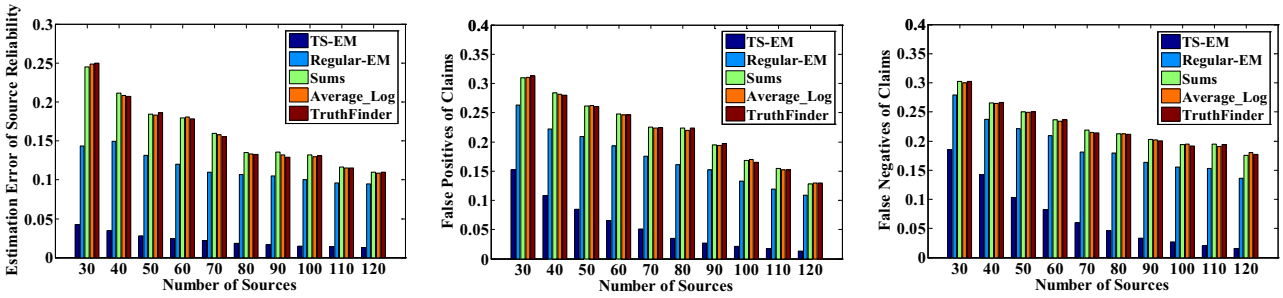
This concludes our general simulations. In the next subsection, we will further evaluate the performance of the TS-EM scheme through a few real-world case studies in a social sensing application.

B. Real Word Case Studies

In this subsection, we evaluate the TS-EM scheme using a real world case study based on Twitter. We choose Twitter as our social sensing application example because it creates an ideal scenario where unreliable content are collected from unvetted data sources with rich time information (e.g., each tweet has its own timestamp) [2]. In our evaluation, we compare *TS-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 explicitly considers the difference in source reliability when it computes the data credibility scores [13]. The third baseline is the *Regular EM*, which was shown to outperform four current truth discovery schemes in social sensing [28].

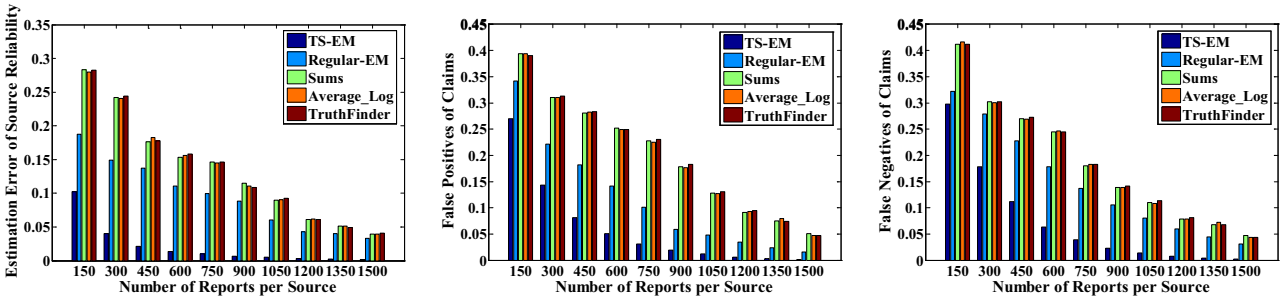
We have implemented the TS-EM scheme and other baselines in Apollo system, a social sensing platform that we have developed to collect tweets from Twitter and track the unfolding of real world events based on the collected tweets [4]. Examples of such events include terrorist attack, hurricane, earthquake, civil unrest and other natural and man-made disasters. Specifically, Apollo has: (i) a data collection front-end that allows users to collect tweets by specifying a set of keywords and/or geo-locations and log the collected tweets; (ii) a data pre-processing component that efficiently clusters similar tweets into the same cluster by using micro-blog data clustering methods [19].

Using the meta-data output by the data pre-processing component of Apollo, 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 Responsiveness Matrix R and Lifespan Vector L. For simplicity, we focused on the binary case here (i.e., $K = 2$, $L = 2$). In particular, we used the following heuristics to categorize the source responsiveness on a report and the lifespan of a claim. First, if the tweet is an original tweet (i.e., not a retweet), it is of high responsiveness. Otherwise it is of low responsiveness. This is based on the observation that the original tweet is generated before the retweets (hence representing a more responsive report of the event). Second, we defined the lifespan of a claim (i.e., cluster) as the difference between the smallest and largest timestamp of tweets in a cluster, which can be easily computed. We then classified the lifespan of each claim into two categories: if it is less than the average lifespan of all claims, it is considered to be short, otherwise it is considered to be long. We note that the above heuristics are only simple approximations to categorize the responsiveness of a source



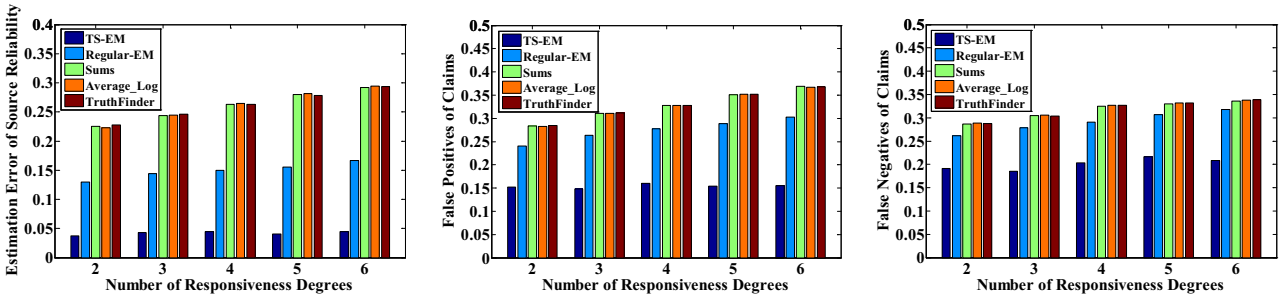
(a) Source Reliability Estimation Accuracy (b) Claims Estimation: False Positives (c) Claims Estimation: False Negatives

Figure 1. Estimation Accuracy versus Number of Sources



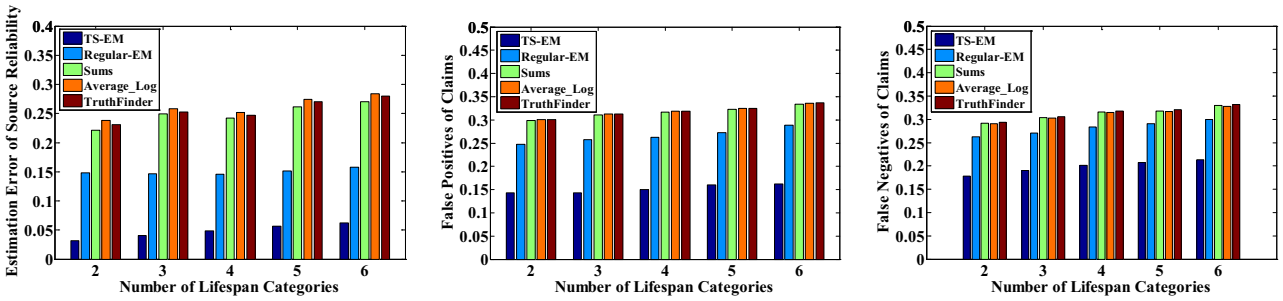
(a) Source Reliability Estimation Accuracy (b) Claims Estimation: False Positives (c) Claims Estimation: False Negatives

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 Number of Source Responsiveness Degrees



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

Figure 4. Estimation Accuracy versus Number of Claim Lifespan Categories

and the lifespan of a claim from real world data. In the future, we will explore more comprehensive techniques to further refine our categorization.

For the purposes of evaluation, we selected a real world data trace from Twitter. The trace was collected by Apollo during the *Paris Charlie Hebdo shooting* event that hap-

pened on January 7, 2015, which caused 12 death and a large scale demonstration in France. There are 74059 tweets contributed by 60984 users in the data trace.

We fed the data trace to the Apollo tool and ran all the compared truth discovery 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.

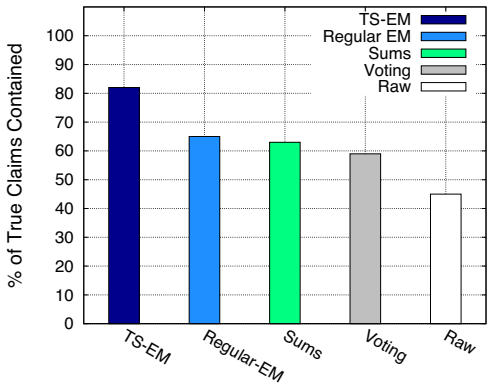


Figure 5. Evaluation on Paris Shooting Trace

Figure 5 shows the result for the Paris Shooting trace. We observe that the TS-EM scheme generally outperforms the Regular EM scheme and other baselines in providing more true claims and suppressing the unconfirmed claims. This is achieved by explicitly incorporating time information (i.e., source responsiveness and claim lifespan) into the maximum likelihood estimation framework. The performance gain of TS-EM scheme compared to Regular EM is significant: *17%*. We also include the reference point called *Raw*, which indicates the average percentage of true claims in a random sample set of raw tweets. We observe that the TS-EM is able to find *37%* more true claims compared to the raw tweets baseline.

VI. Conclusion

This paper develops a time-sensitive maximum likelihood estimation framework to solve the truth discovery problem in social sensing applications. The proposed TS-EM scheme explicitly incorporates the time-sensitive information on both sources and claims (i.e., source responsiveness and claim lifespan) into a rigorous analytical framework. The proposed approach jointly estimates both source reliability and claim correctness using an expectation maximization algorithm. We evaluated the TS-EM scheme through an extensive simulation study and a real world case study in social sensing applications. The results showed TS-EM achieved non-trivial performance gains in improving the truth discovery accuracy compared to the Regular-EM and other state-of-the-art techniques. The results of the paper is important because it lays out an analytical foundation to exploit time-sensitive information in social sensing using a principled approach.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1447795.

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

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

[5] 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.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

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

[8] R. Farmer and B. Glass. *Building web reputation systems*. " O'Reilly Media, Inc.", 2010.

[9] A. Gueziec. Crowd sourced traffic reporting, Apr. 29 2014. US Patent App. 14/265,290.

[10] D. Houser and J. Wooders. Reputation in auctions: Theory, and evidence from eBay. *Journal of Economics & Management Strategy*, 15(2):353–369, 2006.

[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] L. Kaplan, M. Scensoy, and G. de Mel. Trust estimation and fusion of uncertain information by exploiting consistency. In *Information Fusion (FUSION), 2014 17th International Conference on*, pages 1–8. IEEE, 2014.

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

[14] M. Leng and Y.-C. Wu. Low-complexity maximum-likelihood estimator for clock synchronization of wireless sensor nodes under exponential delays. *Signal Processing, IEEE Transactions on*, 59(10):4860–4870, 2011.

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

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

[17] S. S. Pereira, R. Lopez-Valcarce, et al. A diffusion-based em algorithm for distributed estimation in unreliable sensor networks. *Signal Processing Letters, IEEE*, 20(6):595–598, 2013.

[18] Sense Networks. Cab Sense. http://www.cabsense.com.

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

[20] D. Wang, T. Abdelzaher, H. Ahmadi, J. Pasternack, D. Roth, M. Gupta, J. Han, O. Fatemieh, and H. Le. On bayesian interpretation of fact-finding in information networks. In *14th International Conference on Information Fusion (Fusion 2011)*, 2011.

[21] D. Wang, T. Abdelzaher, and L. Kaplan. *Social Sensing: Building Reliable Systems on Unreliable Data*. Morgan Kaufmann, 2015.

[22] 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.

[23] D. Wang, H. Ahmadi, T. Abdelzaher, H. Chenji, R. Stoleru, and C. C. Aggarwal. Optimizing quality-of-information in cost-sensitive sensor data fusion. In *Distributed Computing in Sensor Systems and Workshops (DCOSS), 2011 International Conference on*, pages 1–8. IEEE, 2011.

[24] D. Wang and C. Huang. Confidence-aware truth estimation in social sensing applications. In *The 12th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON 15)*, June 2015.

[25] D. Wang, L. Kaplan, and T. Abdelzaher. Maximum likelihood analysis of conflicting observations in social sensing. *ACM Transactions on Sensor Networks (ToSN)*, Vol. 10, No. 2, Article 30, January, 2014.

[26] 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.

[27] D. Wang, L. Kaplan, T. Abdelzaher, and C. C. Aggarwal. On credibility tradeoffs in assured social sensing. *IEEE Journal On Selected Areas in Communication (JSAC)*, 2013.

[28] 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.

[29] J. Wang, D. Wang, and Y. Zhao. A new reader anti-collision system for rfid. *Journal of Transduction Technology*, 21(8):1411–1416, 2008.

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.