Spatial-Temporal Aware Truth Finding in Big Data Social Sensing Applications

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

***Abstract*—This paper presents a spatial-temporal aware analytical framework to solve the truth finding problem in social sensing applications. Social sensing has emerged as a new big data application paradigm of collecting observations about the physical environment from social sensors (e.g., humans) or devices on their behalf. The collected observations may be true or false, and hence are viewed as binary claims. A fundamental challenge in social sensing applications lies in accurately ascertaining the correctness of claims and the reliability of data sources without knowing either of them a priori. This challenge is referred to as *truth finding*. Significant efforts have been made to address this challenge but two important features were largely missing in the state-of-the-arts solutions: *when* and *where* the claims are reported by a source. In this paper, we develop a new spatial-temporal aware truth finding scheme to explicitly incorporate the *time* information of a claim and *location* information of a source into a rigorous analytical framework. The new truth finding scheme solves a constraint optimization problem to determine both the source reliability and claim correctness. We evaluated the spatial-temporal aware truth finding scheme through both an extensive simulation study and a real world case study using Twitter data feeds. The evaluation results show that our new scheme outperforms all the compared state-of-the-art baselines and significantly improves the truth finding accuracy in social sensing applications.**

***Keywords*-Social Sensing, Spatial-Temporal, Truth Finding, Big Data, Maximum Likelihood Estimation, Expectation Maximization**

I. Introduction

This paper presents a spatial-temporal aware analytical framework to solve the truth finding problem in social sensing applications. We refer social sensing to a new big data application paradigm of collecting observations about the physical environment from social sensors (e.g., humans) or devices on their behalf [20]. This emerging paradigm is enabled by two important technical trends: (i) the proliferation of various sensors in the possession of common individuals (e.g., smartphones); (ii) the popularity of the massive information dissemination opportunities (e.g., online social media). In social sensing, a large crowd of data sources (including both humans or devices they operate) voluntarily contribute massive observations about the physical world in real-time. These observations may be true or false, and hence are viewed as binary *claims*. Examples of social sensing applications include real-time traffic monitoring service using mobile crowdsensing data from driver's smartphones [5], disaster tracking and response service using the publicly available data from online social media (e.g., Twitter, Facebook, Flickr) [28], and geotagging and other location-aware smart city applications using geotagged data from common citizens. Considering the open data contribution paradigm and unvetted nature of data sources, a fundamental challenge 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 finding*.

Previous works have made significant progress to address the truth finding problem in data mining [31], [32], machine learning [11], [12], and networked sensing [19], [21], [22], [24]–[28] communities. The state-of-the-art solutions jointly estimate the reliability of sources and the correctness of their claims based on the observed data (e.g., which claim was reported by which source). However, two key features regarding *when* and *where* the claims are reported were largely missing in the current solutions. *First,* observations about the same claim were assumed to be made at *the same time*. This assumption does not hold in many real world applications. For example, in a geotagging application to find potholes on city streets, it is unlikely that different people will observe and report the same pothole at the exact same time. *Second,* sources were assumed to be at the *same location* when they reported the claims. However, such assumption is too strong in many practical settings. For example, in a disaster response scenario, some sources may report claims about an earthquake at the primary scene of the event while other sources might stay thousands of miles away and make claims based on what they heard from other people. The aforementioned spatial and temporal information has a direct effect on the truth finding results in social sensing.

In this paper, we develop a new spatial-temporal aware truth finding scheme to explicitly incorporate the *time* information of a claim and the *location* information of a source into a rigorous analytical framework. In particular, the new truth finding scheme solves a maximum likelihood estimation (MLE) problem by explicitly considering when and where a claim is made. More specifically, the proposed scheme develops a spatial-temporal aware expectation maxi-

mization (ST-EM) algorithm to statistically assign credibility values to claims and reliability to sources and finds the optimal assignment (in the sense of maximum likelihood estimation) that is most consistent with the observed data as well as the spatial-temporal constraints. We evaluate our spatial-temporal aware truth finding scheme through both an extensive simulation study and a real world case study using Twitter data feeds. The evaluation results demonstrate that our new ST-EM scheme outperforms the state-of-the-art baselines and significantly improves the truth finding accuracy in social sensing.

To summarize, our contributions are as follows:

- To the best of our knowledge, this paper is the first to explicitly consider both spatial and temporal information (i.e., when and where a claim is made) in the truth finding problem in social sensing.
- We develop a rigorous analytical framework that allows us to derive an optimal solution (in the sense of maximum likelihood estimation) for the spatial-temporal aware truth finding problem.
- We show non-trivial performance gains achieved by our ST-EM scheme (i.e., our scheme increased the claim classification precision by 12% compared to state-of-the-art baselines in our real world case study.).

The rest of this paper is organized as follows: we discuss the related work in Section II. The new spatial-temporal truth finding model for social sensing applications is presented in Section III. We develop our maximum likelihood estimation framework and the ST-EM solution 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 application paradigm that empowers common citizens to contribute their daily observations and measurements about the physical world at a very large scale [1]. A comprehensive overview of social sensing applications can be found in [2]. Some early examples of social applications are CenWits [6], CabSense [16], and BikeNet [4]. More recent works have focused on addressing important challenges in social sensing such as energy [10], privacy preservation [3], and incentives design [15]. Since the data collection is open to all, the estimation of data quality and source reliability becomes a critical problem in developing reliable social sensing applications [2]. Some truth finding techniques have been developed to address this problem but they did not explicitly consider *when* and *where* a claim is made by a source [24], [28]. This paper develops a new spatial-temporal aware truth finding scheme that jointly estimates the correctness of claims and the reliability of data sources while explicitly considering the spatial-temporal information obtained from the application.

A good amount of works have been done on the topic of *fact-finding* in data mining and machine learning communities [8], [11], [12], [23], [30]. The *Sums* [8] is one of the earliest fact-finding algorithms that jointly estimates claim credibility and source reliability using an iterative approach. Yin et al. [30] developed TruthFinder that generalized the basic fact-finding scheme by using a pseudo probabilistic model. Pasternack et al. [11] proposed *Investment* and *PooledInvestment* algorithms to explicitly incorporate background knowledge into the fact-finding framework. They further developed a latent credibility analysis (LCA) approach to provide semantic meanings to the analysis results [12]. Other fact-finding techniques also investigate the dependencies between claims and sources respectively [23]. Inspired by the above results, we develop a new spatial-temporal aware maximum likelihood estimation framework to address the truth finding problem in social sensing applications.

Finally, maximum likelihood estimation (MLE) framework has been widely used in the wireless sensor network (WSN) and data fusion communities [9], [13], [17]. For example, Pereira et al. proposed a diffusion-based MLE algorithm for distributed estimation in WSN in the presence of noisy measurements and data faults [13]. Sheng et al. developed a MLE method to infer locations of multiple sources by using acoustic signal energy measurements from individual sensors [17]. Eric et al. deisgned a MLE based approach to aggregate the signals from noisy measurements at remote sensor nodes to a fusion center without any inter-sensor collaborations [9]. However, the above work primarily focused on the estimation of continuous variables from physical sensor measurements. In contrast, this paper focuses on a set of *binary variables* that represent either true or false statements from human sensors. The discrete nature of the estimation variables leads to a more challenging optimization problem that has been solved in this paper.

III. Spatial-Temporal Truth Finding Problem

In this section, we formulate the spatial-temporal truth finding problem in social sensing as a maximum likelihood estimation problem. Consider a social sensing application where a group of M sources, namely, $S_1, S_2, ..., S_M$, who collectively report a set of N claims about the physical world, namely, $C_1, C_2, ..., C_N$. 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 litter locations in a neighborhood, each location may be associated with a claim that is true if the litter is present and false otherwise. In general, any statement about the physical environment, such as "The bridge over X river fell down", "Bombing happened on Y square" and "The building Z 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 litter on the street). 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.

To formulate the spatial-temporal truth finding problem, we need to explicitly consider both the *time* and *location* information when a source reports a claim. First, we define a *Freshness Matrix* R, where the element r_{ij} represents the degree of freshness of C_j when it is reported by S_i. Specifically, r_{ij} is a discrete variable with K different values representing K possible degrees of claim freshness (e.g., fresh, medium, stale). Second, we define a *Location Matrix* G, where the element g_{ij} represents how far S_i is from its reported claim C_j. Specifically, g_{ij} is a discrete variable with L different values representing L distance levels (e.g., nearby, medium, far).

Let us now define a few important terms that we will use in the problem formulation. We denote the *reliability* of source S_i by t_i, which is the probability that a claim is true given that source S_i reports it. Formally, t_i is given by:

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

Considering S_i may report claims at different *time* and *location*, we define $t_i^{k,l}$ as the reliability of S_i when it reports a claim with a freshness degree of k and distance level of l, where $k = 1, ..., K$, $l = 1, ..., L$. Formally, $t_i^{k,l}$ is given by:

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

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

Let us further define $T_i^{k,l}$ to be the (unknown) probability that S_i reports C_j with a freshness degree of k and distance level of l, given that the claim is indeed true. Similarly, we define $F_i^{k,l}$ to be the (unknown) probability that S_i reports C_j with a freshness degree of k and distance level of l, given that the claim is indeed false. Formally, $T_i^{k,l}$ and $F_i^{k,l}$ are:

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

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:

$$\begin{aligned} T_i^{k,l} &= \frac{t_i^{k,l} \times s_i^{k,l}}{d} \\ F_i^{k,l} &= \frac{(1 - t_i^{k,l}) \times s_i^{k,l}}{1 - d} \end{aligned} \tag{5}$$

where d is the prior probability that a randomly chosen claim is true (i.e., $d = p(C_j = 1)$).

Therefore, the spatial-temporal truth finding problem studied in this paper can be formulated as a maximum likelihood estimation (MLE) problem: given the Sensing Matrix SC, the Freshness Matrix R and the Location Matrix G, we aim at estimating the likelihood of the correctness of each claim. Formally, we compute:

$$\forall j, 1 \le j \le N : p(C_j = 1|SC, R, G) \tag{6}$$

IV. A Spatial-Temporal Aware Maximum Likelihood Estimation Approach

A. Background and Mathematical Formulation

In this section, we solve the truth finding problem formulated above by developing a new Spatial-Temporal Aware Expectation-Maximization (ST-EM) algorithm. The EM algorithm is an iterative scheme which alternates between an expectation step (E-step) and a maximization step (M-step) [7]. To apply the EM algorithm in the problem we formulated in Section III, we need to define the observed data X, a vector of unknown estimation parameters θ, a set of unobserved latent variables Z, and a likelihood function $L(\theta; X, Z) = p(X, Z|\theta)$.

The observed data X in our model is the Sensing Matrix SC, Freshness Matrix R and Location Matrix G. The vector of unknown estimation parameters is $\theta = (T_1^{k,l}, T_2^{k,l}, ..., T_M^{k,l}; F_1^{k,l}, F_2^{k,l}, ..., F_M^{k,l}; d)$ $(k = 1, 2, ..., K; l = 1, 2..., L)$, where $T_i^{k,l}$ and $F_i^{k,l}$ are defined in Equation (4). For latent variables, we define a vector of binary variables z_j to indicate whether each claim is true or false (i.e., $z_j = 1$ if C_j is true and 0 otherwise). Furthermore, in order to explicitly incorporate the spatial-temporal information into our MLE framework, we also define two set of binary indicators r_{ij}^k and g_{ij}^l, where $r_{ij}^k = 1$ when $r_{ij} = k$ and $r_{ij}^k = 0$ otherwise. Similarly, $g_{ij}^l = 1$ when $g_{ij} = l$ and $g_{ij}^l = 0$ otherwise. Using the above definitions, we can write the likelihood function of our problem as follows:

$$\begin{aligned} &L(\theta; X, Z) = p(X, Z|\theta) \\ &= \prod_{j=1}^{N}\Bigg\{\prod_{i=1}^{M}\Bigg[\prod_{k=1}^{K}\{\prod_{l=1}^{L}(T_i^{k,l})^{S_iC_j \ \&\& \ r_{ij}^k \ \&\& \ g_{ij}^l} \\ &\times (1 - \sum_{k=1}^{K}\sum_{l=1}^{L} T_i^{k,l})^{(1-S_iC_j)} \times d \times z_j\}\Bigg] \\ &+ \prod_{i=1}^{M}\Bigg[\prod_{k=1}^{K}\{\prod_{l=1}^{L}(F_i^{k,l})^{S_iC_j \ \&\& \ r_{ij}^k \ \&\& \ g_{ij}^l} \\ &\times (1 - \sum_{k=1}^{K}\sum_{l=1}^{L} F_i^{k,l})^{(1-S_iC_j)} \times (1 - d) \times (1 - z_j)\}\Bigg]\Bigg\} \end{aligned} \tag{7}$$

where the "&&" represents the "AND" logic for binary variables. The likelihood function represents the likelihood of the observed data X and the values of hidden variables Z given the estimation parameters θ.

B. Deriving the E and M Steps

Given the likelihood function, we can derive E and M steps of the proposed spatial-temporal aware EM scheme. First, we derive the Q function in E-step as follows:

$$\begin{aligned}
&Q(\theta|\theta^{(n)}) = E_{Z|X,\theta^{(n)}}[logL(\theta; X, Z)] \\
&= \sum_{j=1}^{N} p(z_j = 1|X_j, \theta^{(n)}) \\
&\times \sum_{i=1}^{M} \Bigg[\sum_{k=1}^{K}\sum_{l=1}^{L} (S_iC_j \ \&\& \ r_{ij}^k \ \&\& \ g_{ij}^l) \times logT_i^{k,l} \\
&+ (1 - S_iC_j) \times log(1 - \sum_{k=1}^{K}\sum_{l=1}^{L} T_i^{k,l}) + logd \Bigg] \\
&+ p(z_j = 0|X_j, \theta^{(n)}) \\
&\times \sum_{i=1}^{M} \Bigg[\sum_{k=1}^{K}\sum_{l=1}^{L} (S_iC_j \ \&\& \ r_{ij}^k \ \&\& \ g_{ij}^l) \times logF_i^{k,l} \\
&+ (1 - S_iC_j) \times log(1 - \sum_{k=1}^{K}\sum_{l=1}^{L} F_i^{k,l}) + log(1 - d) \Bigg] \Bigg\}
\end{aligned} \tag{8}$$

Then we define $Z(n,j) = p(z_j = 1|X_j, \theta^{(n)})$. It is the conditional probability of the claim C_j to be true given the observed data X_j and current estimate of θ. $Z(n,j)$ can be further expressed as:

$$\begin{aligned}
Z(n,j) &= p(z_j = 1|X_j, \theta^{(n)}) \\
&= \frac{p(z_j = 1; X_j, \theta^{(n)})}{p(X_j, \theta^{(n)})} \\
&= \frac{A(n,j) \times (d)^{(n)}}{A(n,j) \times (d)^{(n)} + B(n,j) \times (1 - (d)^{(n)})}
\end{aligned} \tag{9}$$

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

$$\begin{aligned}
A(n,j) &= p(X_j, \theta^{(n)}|z_j = 1) \\
&= \prod_{i=1}^{M} \Big\{ \prod_{k=1}^{K}\prod_{l=1}^{L} (T_i^{k,l})^{S_iC_j \ \&\& \ r_{ij}^k \ \&\& \ g_{ij}^l} \\
&\times (1 - \sum_{k=1}^{K}\sum_{l=1}^{L} T_i^{k,l})^{(1 - S_iC_j)} \Big\}
\end{aligned} \tag{10}$$

$$\begin{aligned}
B(n,j) &= p(X_j, \theta^{(n)}|z_j = 0) \\
&= \prod_{i=1}^{M} \Big\{ \prod_{k=1}^{K}\prod_{l=1}^{L} (F_i^{k,l})^{S_iC_j \ \&\& \ r_{ij}^k \ \&\& \ g_{ij}^l} \\
&\times (1 - \sum_{k=1}^{K}\sum_{l=1}^{L} F_i^{k,l})^{(1 - S_iC_j)} \Big\}
\end{aligned} \tag{11}$$

$A(n,j)$ and $B(n,j)$ represent conditional probability of observed reports about C_j and current estimation of unknown parameter θ, given that claim C_j is true and false respectively.

Next we simplify Equation (8) by replacing the conditional probability of $p(z_j = 1|X_j, \theta^{(n)})$ with $Z(n,j)$.

$$\begin{aligned}
&Q(\theta|\theta^{(n)}) = E_{Z|X,\theta^{(n)}}[logL(\theta; X, Z)] \\
&= \sum_{j=1}^{N} Z(n,j) \\
&\times \sum_{i=1}^{M} \Bigg[\sum_{k=1}^{K}\sum_{l=1}^{L} (S_iC_j \ \&\& \ r_{ij}^k \ \&\& \ g_{ij}^l) \times logT_i^{k,l} \\
&+ (1 - S_iC_j) \times log(1 - \sum_{k=1}^{K}\sum_{l=1}^{L} T_i^{k,l}) + logd \Bigg] \\
&+ (1 - Z(n,j)) \\
&\times \sum_{i=1}^{M} \Bigg[\sum_{k=1}^{K}\sum_{l=1}^{L} (S_iC_j \ \&\& \ r_{ij}^k \ \&\& \ g_{ij}^l) \times logF_i^{k,l} \\
&+ (1 - S_iC_j) \times log(1 - \sum_{k=1}^{K}\sum_{l=1}^{L} F_i^{k,l}) + log(1 - d) \Bigg] \Bigg\}
\end{aligned} \tag{12}$$

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 (8) 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 $\frac{\partial Q}{\partial d} = 0$. The optimal $(T_i^{k,l})^*$, $(F_i^{k,l})^*$ and d^* are given as follows:

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

where $Z(n,j) = p(z_j = 1|X_j, \theta^{(n)})$ and $SW_i^{k,l}$ is the set of claims that source S_i reports with a freshness degree of k and distance level of l. N is the total number of claims in the Sensing Matrix SC.

C. The Spatial-Temporal Aware EM Algorithm

Algorithm 1 Spatial-Temporal Aware EM Algorithm

Initialize θ ($T_i^{k,l} = s_i^{k,l}$, $F_i^{k,l} = 0.5 \times s_i^{k,l}$, 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 (9)
  end for
  $\theta^{(n+1)} = \theta^{(n)}$
  for $i = 1 : M$ **do**
    compute $(T_i^{k,l})^{(n+1)}, (F_i^{k,l})^{(n+1)}, d^{(n+1)}$ based on Equation (13)
    update $(T_i^{k,l})^{(n)}$, $(F_i^{k,l})^{(n)}$, $d^{(n)}$ with $(T_i^{k,l})^{(n+1)}$, $(F_i^{k,l})^{(n+1)}$, $d^{(n+1)}$ in $\theta^{(n+1)}$
  end for
  $n = n + 1$
end while
Let $(Z_j)^c$ = converged value of $Z(n, j)$
Let $(T_i^{k,l})^c$ = converged value of $(T_i^{k,l})^{(n)}$;
Let $(F_i^{k,l})^c$ = converged value of $(F_i^{k,l})^{(n)}$;
Let 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^l is false
  end if
end for
for $i = 1 : M$ **do**
  calculate $(t_i^{k,l})^*$ from $(T_i^{k,l})^c$, $(F_i^{k,l})^c$ and d^c based on Equation (5)
  calculate $t_i{}^*$ form $(t_i^{k,l})^*$ based on Equation (3)
end for
Return the MLE on source reliability $t_i{}^*$ and corresponding judgment on the correctness of claim C_j.

In summary, the input of the ST-EM algorithm is the Sensing Matrix SC, Freshness Matrix R and Location Matrix G obtained from the social sensing data. The output is the maximum likelihood estimation of estimation parameters (i.e.,θ) and hidden variables (i.e., Z). The estimation results can be used to compute both source reliability and the correctness of claims. More specifically, the E-step and M-step of the ST-EM algorithm are shown in Equation (9) and Equation (13) respectively. The convergence analysis has been done for EM scheme and it is beyond the scope of this paper [29]. We summarize the ST-EM scheme in Algorithm 1.

V. Evaluation

In this section, we conduct experiments to evaluate the proposed spatial-temporal aware EM scheme (ST-EM) through both a simulation study and a real world case study. The experimental results show that significant performance improvements can be achieved by the proposed ST-EM scheme compared to other state-of-the-art algorithms.

A. Simulation Study

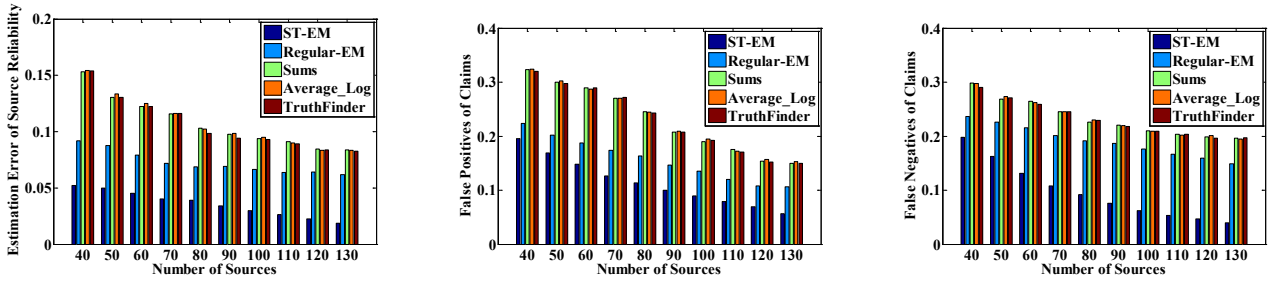
In this subsection, we evaluate the performance of the proposed ST-EM algorithm through an extensive simulation study. In particular, we evaluated two performance metrics: (i) the accuracy of source reliability estimation, (ii) the false positives and false negatives of claim classifications. We built a social sensing simulator in Python 2.7 and compared the proposed scheme (i.e., ST-EM) with the Regular-EM [28] and other three state-of-the-art baselines: Sums [8], Average_Log [11] and TruthFinder [31]. In the simulation, we generate a random number of sources and claims. Each source S_i is associated with two parameters: (i) a random reliability t_i that represents the ground truth probability that source S_i reports correct claims; (ii) the probability of source S_i reporting claims (i.e., s_i). For the claims, each claim C_j represents a binary statement of a physical event (i.e., $True$ or $False$). Note that source S_i may report claim C_j with a certain degree of freshness k, $k = 1, ..., K$ at a certain distance level l, $l = 1, ..., L$. For the following experiments, we set $K = 3$ and $L = 3$ and evaluate the performance of ST-EM by varying different parameters of the model. The reported results are averaged over 100 experiments.

In the first experiment, we evaluated the performance of ST-EM and other baselines while varying the number of sources in this system. The total number of claims was fixed at 2000, of which 1000 was true and 1000 was false. We set the average number of reports per source to 200. The reports are uniformly distributed across different freshness degrees and different distance levels. In this experiment, we vary the number of sources from 40 to 130. Results are reported in Figure 1. We observe that ST-EM achieves the smallest error on source reliability estimation and the least false positives and false negatives on claim classification when the number of sources changes in the system.

In the second experiment, we evaluated the performance of ST-EM and other baselines by changing the average number of reports made per source. In this experiment, we set the number of sources to 50. We change the average number of reports made per source from 100 to 1000. Other experiment parameters are kept the same as the first experiment. Results are shown in Figure 2. We observe that ST-EM outperforms all baselines in terms of source reliability estimations as well as the false positives and false negatives of the claim classification. We can also note that the performance gain of ST-EM is significant as the average number of reports made per source changes in our system.

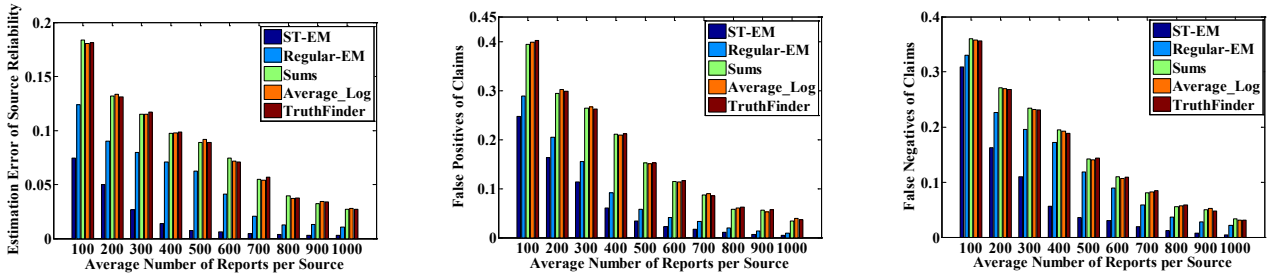
In the third experiment, we evaluated the performance of all schemes when the fraction of fresh reports varies. In this experiment, we changed the fraction of fresh reports from 0.1 to 0.9. We set fractions of reports made with different distance levels to be equal. The number of sources was set to 50 and the average number of reports made per source was set to 300. We keep other experiment configurations the same as before. Results are shown in Figure 3. Observe that ST-EM scheme consistently outperforms all compared baselines when the fraction of the fresh reports changes in the system.

In the last experiment, we evaluate the performance of all



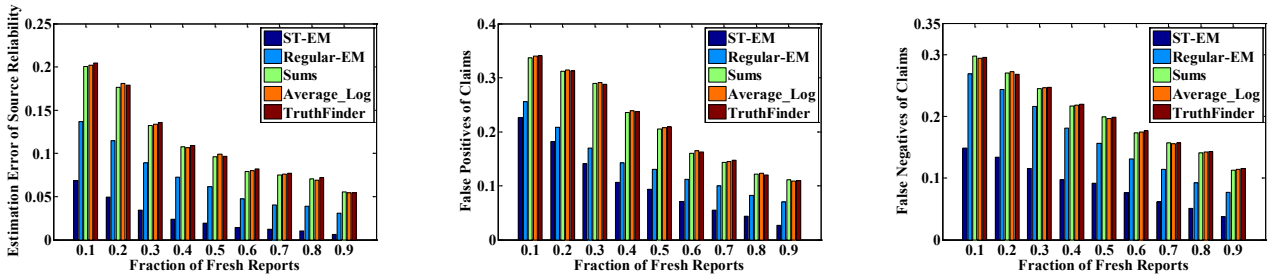
(a) Source Reliability Estimation Accuracy (b) Claims Estimation: False Positives (c) Claims Estimation: False Negatives

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

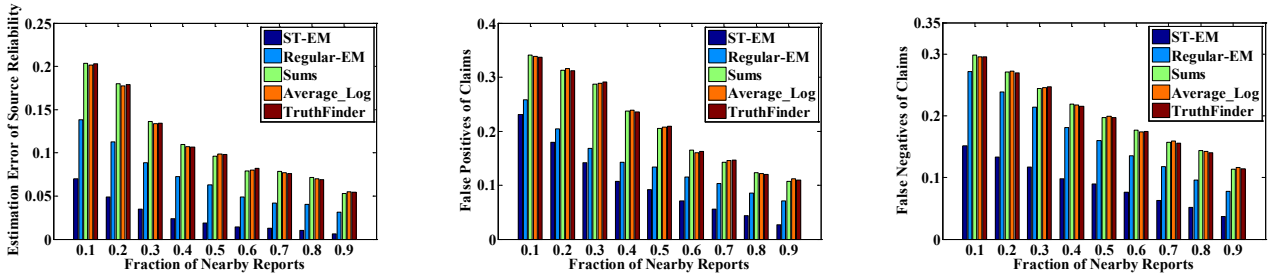
Figure 3. Estimation Accuracy versus Fraction of Fresh Reports

schemes when the fraction of reports made from a nearby location of the event (i.e., nearby reports) changes. In this experiment, we varied the fraction of nearby reports from 0.1 to 0.9. We set the fractions of reports with different degrees of freshness to be equal. Other experiment parameters are also kept the same as the third experiment. Results are reported in Figure 4. We observe that ST-EM also achieves the best performance compared with other baselines when the fraction of nearby reports changes in the system.

B. A Real Word Case Study

In this subsection, we evaluate the ST-EM scheme using a real-world case study based on Twitter, which provides a popular platform for average people to share massive unvetted observations in real-time. In our evaluation, we compare the performance of the proposed ST-EM algorithm with three baselines from current literature. The first one is the *Regular EM* proposed for participatory sensing applications in [28]. The second baseline is *Voting*, where claims correctness is estimated by the number of times the same tweet is reported on Twitter. The third baseline is *Sums* [8] which takes into account the differences of sources reliability when estimating the correctness of claims. For the purpose of evaluation, we collected a Twitter dataset during *Boston Marathon bombing* event that happened on April 15, 2013 and subsequent shootings and manhunt events. In particular, the dataset consists of 123,402 tweets and 101,209 users. We first used a micro-blog based clustering algorithm [18] to cluster similar tweets into the same cluster. Then we generate 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 Freshness Matrix R and Location Matrix G. For simplicity, we focus on the binary case for R and trinary cases for G (i.e., $K = 2$,$L = 3$). In



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

Figure 4. Estimation Accuracy versus Fraction of Nearby Reports

particular, we use the following heuristics: (i) we classify the freshness of a tweet into two categories: if the tweet is an original tweet (i.e., not a retweet), it is fresh. Otherwise, it is non-fresh. This is motivated by the observation that the create time of an original tweet is earlier than that of retweets. (ii) We used the geo-coordinates field embedded in a tweet to compute the distance between a source's location and the center of the event (i.e., d_l). We then compared the d_l with some selected threshold (e.g., we used 100 miles as a threshold to cover the major area of Boston in our experiment). The source distance level can be classified into three categories based on the comparison results: nearby (if $d_l \leq threshold$), far (if $d_l > threshold$), undetermined (the tweet does not contain the geo-coordinates). Note that the above heuristics are only the first approximation to categorize the time and location information of a report from real world data. In the future, we will explore more comprehensive techniques to further refine our categorization.

We evaluated all the compared schemes based on the Twitter data feeds. The output of these schemes was manually graded to determine the credibility of claims. Due to manpower limitations, we manually graded the top 50 claims ranked 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 satisfy the requirement of true claims.

We note that some of the unconfirmed claims may be possibly true but cannot be independently verified through external sources. Hence, our evaluation provides *pessimistic* performance bounds on the estimates by taking all the unconfirmed claims as false.

The results are shown in Figure 5. We observe that ST-EM outperforms the Regular EM scheme and other baselines in providing more true claims and suppressing the unconfirmed claims. In particular, ST-EM scheme found 12% more true claims than the best performed baseline (i.e., Regular EM). This performance improvement is achieved

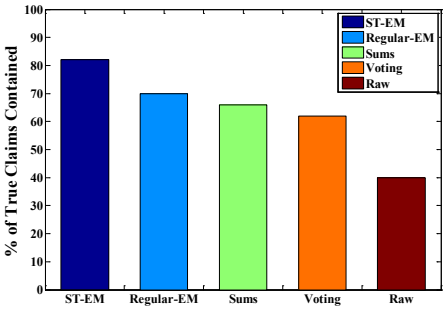


Figure 5. Evaluation on Twitter Data Feeds from Boston Bombing Trace

by explicitly incorporating the spatio-temporal information into the maximum likelihood estimation framework. We also included the reference point called *Raw*, which indicates the average percentage of true claims in a random sample set of raw tweets.

VI. Conclusion

This paper presents a spatio-temporal maximum likelihood estimation framework to solve the truth discovery problem in social sensing applications. The proposed ST-EM scheme explicitly incorporates the spatio-temporal information into a rigorous analytical framework. We evaluated the ST-EM scheme through an extensive simulation study and a real world case study based on Twitter. The results showed ST-EM achieved good performance gains 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 explore spatio-temporal information to solve truth finding problem in social sensing applications.

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] 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.

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

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

[6] 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.

[7] J.Han, M.Kamber, and J. Pei. *Data Mining: Concepts and Techniques, Third Edition*. Morgan Kaufman, 2011.

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

[9] E. J. Msechu and G. B. Giannakis. Sensor-centric data reduction for estimation with wsns via censoring and quantization. *Signal Processing, IEEE Transactions on*, 60(1):400–414, 2012.

[10] 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.

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

[12] 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.

[13] 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.

[14] S. Reddy, D. Estrin, M. Hansen, and M. Srivastava. Examining micro-payments for participatory sensing data collections. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 33–36. ACM, 2010.

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

[16] 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.

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

[18] D. Wang, T. Abdelzaher, and L. Kaplan. Surrogate mobile sensing. *Communications Magazine, IEEE*, 52(8):36–41, 2014.

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

[20] D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal. On quantifying the accuracy of maximum likelihood estimation of participant reliability in social sensing. In *DMSN11: 8th International Workshop on Data Management for Sensor Networks*, August 2011.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

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

[31] 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.