Exploitation of Physical Constraints for Reliable Social Sensing

Dong Wang[1], Tarek Abdelzaher[1], Lance Kaplan[2], Raghu Ganti[3], Shaohan Hu[1], Hengchang Liu[1,4]
[1]Department of Computer Science, University of Illinois at Urbana Champaign, Urbana, IL 61801
[2]Networked Sensing and Fusion Branch, US Army Research Labs, Adelphi, MD 20783
[3]IBM Research, Yorktown Heights, NY 10598
[4]Department of Computer Science, University of Science and Technology of China, Hefei, Anhui 230027

***Abstract*—This paper develops and evaluates algorithms for exploiting physical constraints to improve the reliability of social sensing. Social sensing refers to applications where a group of sources (e.g., individuals and their mobile devices) volunteer to collect observations about the physical world. A key challenge in social sensing is that the reliability of sources and their devices is generally unknown, which makes it non-trivial to assess the correctness of collected observations. To solve this problem, the paper adopts a cyber-physical approach, where assessment of correctness of individual observations is aided by knowledge of physical constraints on both sources and observed variables to compensate for the lack of information on source reliability. We cast the problem as one of maximum likelihood estimation. The goal is to jointly estimate both (i) the latent physical state of the observed environment, and (ii) the inferred reliability of individual sources such that they are maximally consistent with both provenance information (who claimed what) and physical constraints. We evaluate the new framework through a real-world social sensing application. The results demonstrate significant performance gains in estimation accuracy of both source reliability and observation correctness.**

I. Introduction

This paper investigates the exploitation of physical constraints to improve the reliability of social sensing applications. We refer by *social sensing* to a broad set of applications, where sources, such as humans and digital devices they operate, collect information about the physical world for purposes of mutual interest. In social sensing, humans can play different roles by acting as sensor carriers [21] (e.g., opportunistic sensing), sensor operators [4] (e.g., participatory sensing) or sensor themselves [36]. The proliferation of mobile devices with sensors, such as smart phones, has significantly increased the popularity of social sensing. Examples of recent applications include optimization of daily commute [18], [44], reduction of carbon footprint [10], [20], disaster response [17], [33] and pollution monitoring [24], [28], to name a few. Due to the inclusive nature of data collection in social sensing (i.e., anyone can participate) and the unknown reliability of information sources, much recent work focused on estimating the likelihood of correctness of collected data [26], [36], [42].

The novelty of this work comes from adopting a cyber-physical approach to the problem of assessing correctness of collected data, wherein physical constraints are exploited to compensate for unknown source reliability. We consider two types of constraints; namely, (i) source constraints that, combined with source location information, offer an understanding of what individual sources observed, and (ii) constraints on the observed variables themselves that arise when these variables are not independent. Together, these constraints shape the likelihood function that quantifies the odds of the observations at hand. We then maximize the resulting likelihood function with respect to hypotheses on the correctness of individual observations as well as hypotheses on the reliability of individual sources. We show that the maximum likelihood estimate thus obtained is a lot more accurate than one that does not take physical constraints into account.

The use of a maximum likelihood estimation framework to jointly compute both the reliability of sources in social sensing applications and the correctness of the data they report was recently described by the authors [36], but without taking physical constraints into account. The advantage of maximum-likelihood estimation lies in the feasibility of computing rigorous estimation accuracy bounds [35], hence not only arriving at the top hypothesis, but also quantifying how good it is. The main contribution of this paper lies in developing the analytical foundation for exploiting physical constraints within the aforementioned maximum likelihood estimation framework. A physically-aware Expectation Maximization (EM) algorithm is developed that is empirically shown to converge to a more accurate solution than the above baseline, thanks to taking the constraints into account.

Our work is related to machine learning literature on constrained conditional models [5], [26]. Unlike that literature, we do not limit our approach to simple linear models [5] nor require that dependencies and constraints be deterministic [26]. Instead, the framework developed in this paper is general enough to (i) solve the optimization problem for *non-linear* models abstracted from social sensing applications with physical constraints (as shown in Section III and IV), and (ii) incorporate *probabilistic* dependencies.

Finally, contrary to work that focuses on maximum-likelihood estimation of continuous variables given continuous models of physical phenomena, which appears in both sensor networks and data fusion literature [3], [23], [39], we focus on estimating discrete variables. Specifically, we estimate the values of a string of generally non-independent Booleans that can either be true or false. The discrete nature of the estimated variables makes our optimization problem harder, as it gives rise to an integer programming problem whose solution space increases exponentially. We show that the complexity of our results critically depends on the number of variables that appear in an *individual constraint*, as opposed to the number of variables in the system. Hence, the approach scales well to large numbers of estimated variables as long as constraints are localized. We evaluate the scheme through a real-world social sensing application. Results show significant performance improvements in both source reliability estimation and claim correctness estimation, achieved by incorporating physical information into the estimation framework.

The rest of the paper is organized as follows. Section II formulates the problem of reliable social sensing, where the goal is to estimate correctness of claims. Section III and Section IV solve the problem while leveraging source constraints and observed-variable constraints, respectively. Evaluation results are presented in Section V. We review the related work in Section VII. Finally, we conclude the paper in Section VIII.

II. The Problem Formulation

Much prior research in sensor networks [3], [39] and estimation theory [14], [29] considered filtering observations of continuous variables in a maximum-likelihood fashion to separate signal from noise. While continuous variables are common in sensing, an important subset of sensing applications deals primarily with discrete (and especially binary) variables. Interestingly, noise reduction in the case of binary variables is more challenging, because discretization gives rise to likelihood functions that are not continuous, hence leading to integer programming problems, known to be NP-complete.

Binary variables arise in many applications where the state of the physical environment can be represented by a set of statements, each is either true or false. For example, in an application where the goal is to find free parking spots around campus, each legal parking spot may be associated with one variable that is true if the spot is available and false otherwise. Similarly, in an application that reports offensive graffiti on campus walls, each location may be associated with a variable that is true if offensive graffiti is present and false otherwise. In general, any statement about the physical world, such as "Main Street is flooded", "The airport is closed", or "The suspect was seen on Elm Street" can be thought of as a binary variable that is true if the statement is correct, and false if it is not.

Accordingly, in this paper, we consider social sensing applications, where a group of M sources, $S_1, ..., S_M$, observe a set of N binary variables, $C_1, ..., C_N$. Each variable C_j is associated with a location, L_j. Sources report some of their observations. We call a reported observation a *claim*. Please note that observations and claims are *not* used interchangeably in this paper. Observations refer to what a source had an opportunity to witness. Claims refer to what the source reported that they witnessed. We use the word *claim* to emphasize that we do not, in general, know whether the report is correct or not. We assume, without loss of generality, that the "normal" state of each variable is negative (e.g., no free parking spots and no graffiti on walls). Hence, sources report only when a positive value is encountered. As mentioned above, the reliability of individual sources is not known. In other words, we do not know the "noise model" that determines the odds that a source reports incorrectly.

The contribution of this paper lies in exploiting physical constraints to compensate for the lack of information on source reliability. Two types of physical constraints are exploited:

- *Constraints on sources:* A source constraint simply states that a source can only observe co-located physical variables. In other words, it can only report C_j if it visited location L_j. The granularity of locations is application specific. However, given location granularity in a particular application context, this constraint allows us to understand which variables a source had an opportunity to observe. Hence, for example, when a source does not report an event that others claim they observed, we can tell whether or not the silence should decrease our confidence in the reported observation, depending on whether or not the silent source was co-located with the alleged event.
- *Constraints on observed variables:* We exploit the fact that observed variables may be correlated, which can be expressed by a joint probability distribution on the underlying variables. For example, traffic speed at different locations of the same freeway may be related by a joint probability distribution that favors similar speeds. This probabilistic knowledge gives us a basis for assessing how internally consistent a set of reported observations is.

Let S_i represent the i^{th} source and C_j represent the j^{th} variable. We say that S_i *observed* C_j if the source visited location L_j. We say that a source S_i made a *claim* S_iC_j if the source reported that C_j was true. We generically denote by $P(C_j = 1|x)$ and $P(C_j = 0|x)$ the conditional probability that variable C_j is indeed true or false, given x, respectively. We denote by t_i the (unknown) probability that a claim is correct given that source S_i reported it, $t_i = P(C_j = 1|S_iC_j)$. Different sources may make different numbers of claims. The probability that source S_i makes a claim is s_i. Formally, $s_i = P(S_iC_j|S_i \text{ observes } C_j)$.

We further define a_i to be the (unknown) probability that source S_i correctly reports a claim given that the underlying variable is indeed true and the source observed it. Similarly, we denote by b_i the (unknown) probability that source S_i falsely reports a claim when the underlying variable is in reality false and the source observed it. More formally:

$$\begin{aligned} a_i &= P(S_iC_j|C_j = 1, S_i \text{ observes } C_j) \\ b_i &= P(S_iC_j|C_j = 0, S_i \text{ observes } C_j) \end{aligned} \tag{1}$$

From the definitions above, we can determine the following relationships using the Bayesian theorem:

$$\begin{aligned} a_i &= P(S_iC_j|C_j = 1, S_i \text{ observes } C_j) \\ &= \frac{P(C_j = 1|S_iC_j, S_i \text{ observes } C_j)P(S_iC_j|S_i \text{ observes } C_j)}{P(C_j = 1|S_i \text{ observes } C_j)} \\ b_i &= P(S_iC_j|C_j = 0, S_i \text{ observes } C_j) \\ &= \frac{P(C_j = 0|S_iC_j, S_i \text{ observes } C_j)P(S_iC_j|S_i \text{ observes } C_j)}{P(C_j = 0|S_i \text{ observes } C_j)} \end{aligned} \tag{2}$$

We also define d_i to be the (unknown) probability $P(C_j = 1|S_i \text{ observes } C_j)$. It should be noted that it does depend on variable j. This is the proportion of variables that source S_i observes that happen to be true. Note that, the probability that a source makes a claim is proportional to the number of claims made by the source over the total number of variables observed by the source. Plugging these, together with t_i, into the definition of a_i and b_i, given in Equation (2), we get the relationship between the terms we defined above:

$$a_i = \frac{t_i \times s_i}{d_i} \qquad b_i = \frac{(1-t_i) \times s_i}{1-d_i}$$
$$d_i = P(C_j = 1) \quad j \in \mathcal{C}_i \tag{3}$$

where $\mathcal{C}_i$ is the set of variables that S_i observed. The input to our algorithm is: (i) the *claim matrix* SC, where $S_iC_j = 1$ when source S_i reports that C_j is true, and $S_iC_j = 0$ otherwise; and (ii) the source's opportunities to observe represented by a *knowledge matrix* SK, where $S_iK_j = 1$ when source S_i has the opportunity to observe C_j and $S_iK_j = 0$ otherwise. The output of the algorithm is the probability that variable C_j is true, for each j and the reliability t_i of source S_i, for each i. More formally:

$$\forall j, 1 \leq j \leq N : P(C_j = 1 | SC, SK)$$
$$\forall i, 1 \leq i \leq M : P(C_j = 1 | S_iC_j) \tag{4}$$

To account for non-independence among the observed variables, we further denote the set of all such constraints (expressed as joint distributions of dependent variables) by JD. The inputs to the algorithm become the SC, SK matrices and the set JD of constraints (joint distributions), mentioned above. The output is:

$$\forall j, 1 \leq j \leq N : P(C_j = 1 | SC, SK, JD)$$
$$\forall i, 1 \leq i \leq M : P(C_j = 1 | S_iC_j) \tag{5}$$

Below, we solve the aforementioned problems using the expectation maximization (EM) algorithm. EM [6] is a general algorithm for finding the maximum likelihood estimates of parameters in a statistic model, where the likelihood function involves latent variables. Applying EM requires formulating the likelihood function, $L(\theta; X, Z) = p(X, Z|\theta)$, where θ is the estimated parameter vector, X is the observed data, and Z is the latent variables vector. The algorithm then maximizes likelihood iteratively by alternating between two steps:

- E-step: Compute the expected log likelihood function, where the expectation is taken with respect to the computed conditional distribution of the latent variables given the current settings and observed data.

$$Q\left(\theta|\theta^{(t)}\right) = E_{Z|X,\theta^{(t)}}[\log L(\theta; X, Z)] \tag{6}$$

- M-step: Find the parameters that maximize the Q function in the E-step to be used as the estimate of θ for the next iteration.

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q\left(\theta|\theta^{(t)}\right) \tag{7}$$

Following the approach described in our previous work [36], we define a latent variable z_j to denote our estimated value of variable C_j, for each j (indicating whether it is true or not). Initially, we set $p(z_j = 1) = d_j$. This constitutes the latent vector Z above. We further define X to be the claim matrix SC, where X_j represents the j^{th} column of the SC matrix (i.e., claims of the j^{th} variable by all sources). The parameter vector we want to estimate is $\theta = (a_1, a_2, ...a_M; b_1, b_2, ...b_M; d_1, d_2, ..., d_N)$. In the following sections, we incorporate the physical constraints into the above model, which is the new contribution of the paper.

III. Accounting for Opportunity to Observe

In this section, we incorporate the source constraints into the Expectation-Maximization (EM) algorithm. We call this EM scheme, EM with *opportunity to observe* (OtO EM).

A. Deriving the Likelihood

When we consider source constraints in the likelihood function, we assume sources only claim variables they observe, and hence the probability of a source claiming a variable he/she does not have an opportunity to observe is 0. For simplicity, we first assume that all variables are independent, then relax this assumption later in Section IV. Under these assumptions, the new likelihood function $L(\theta; X, Z)$ that incorporates the source constraints is given by:

$$\begin{aligned} L(\theta; X, Z) &= p(X, Z|\theta) \\ &= \prod_{j=1}^{N} p(z_j) \times p(X_j | z_j, \theta) \\ &= \prod_{j=1}^{N} \prod_{i \in \mathcal{S}_j} p(z_j) \times \alpha_{i,j} \\ &\text{where } \mathcal{S}_j : \text{Set of sources observed } C_j \end{aligned} \tag{8}$$

where

$$p(z_j) = \begin{cases} d_j & z_j = 1 \\ (1 - d_j) & z_j = 0 \end{cases}$$
$$\alpha_{i,j} = \begin{cases} a_i & z_j = 1, S_iC_j = 1 \\ (1 - a_i) & z_j = 1, S_iC_j = 0 \\ b_i & z_j = 0, S_iC_j = 1 \\ (1 - b_i) & z_j = 0, S_iC_j = 0 \end{cases} \tag{9}$$

Note that, in the likelihood function, we only consider the probability contribution from sources who actually *observe* a variable (e.g., $i \in \mathcal{S}_j$ for C_j). This is an important change from our previous framework [36]. This change allows us to nicely incorporate the source constraints (name, source opportunity to observe) into the maximum likelihood estimation framework.

Using the above likelihood function, we can derive the corresponding E-Step and M-Step of OtO EM scheme. The detailed derivations are shown in Section IX-A.

B. The OtO EM Algorithm

In summary, the inputs to the OtO EM algorithm are (i) the claim matrix SC from social sensing data and (ii) the knowledge matrix SK describing the *source constraints*. The output is the maximum likelihood estimate of source reliability and the probability of claim correctness. Compared to the regular EM algorithm we derived in our previous work [36], we provided source constraints as a new input into the framework and imposed them on the E-step and M-step. Our algorithm begins by initializing the parameter θ with random values between 0 and 1. The algorithm then performs the new derived E-steps and M-steps iteratively until θ converges. Convergence analysis for EM was studied in literature and is out of the

Algorithm 1 Expectation Maximization Algorithm with Source Constraints (OtO EM)

Initialize θ with random values between 0 and 1
while $\theta^{(t)}$ does not converge **do**
    for $j = 1 : N$ **do**
        compute $Z(t, j)$ based on Equation (12)
    end for
    $\theta^{(t+1)} = \theta^{(t)}$
    for $i = 1 : M$ **do**
        compute $a_i^{(t+1)}, b_i^{(t+1)}, d_j^{(t+1)}$ based on Equation (13)
        update $a_i^{(t)}, b_i^{(t)}, d_j^{(t)}$ with $a_i^{(t+1)}, b_i^{(t+1)}, d_j^{(t+1)}$ in $\theta^{(t+1)}$
    end for
    $t = t + 1$
end while
Let Z_j^c = converged value of $Z(t, j)$
Let a_i^c = converged value of $a_i^{(t)}$; b_i^c = converged value of $b_i^{(t)}$; d_i^c = converged value of $d_j^{(t)}$ $\quad j \in \mathcal{C}_i$
for $j = 1 : N$ **do**
    if $Z_j^c \geq threshold$ **then**
        C_j is true
    else
        C_j is false
    end if
end for
for $i = 1 : M$ **do**
    calculate t_i^* from a_i^c, b_i^c and d_i^c
end for
Return the classification on variables and reliability estimation of sources

scope for this paper [40].[1] Since each observed variable is binary, we can classify variables as either true or false based on the converged value of $Z(t, j)$. Specifically, C_j is considered true if Z_j^c goes beyond some threshold (e.g., 0.5) and false otherwise. We can also compute the estimated t_i of each source from the converged values of $\theta^{(t)}$ (i.e., a_i^c, b_i^c and d_i^c) based on Equation (3). Algorithm 1 shows the pseudocode of OtO EM.

IV. Accounting for Dependency Constraints

In this section, we derive an EM scheme that considers constraints on observed variables. We call this EM scheme, EM with *dependent variables* (DV EM). For clarity, we first ignore the source constraints derived in the previous section (i.e., assume that each source observes all variables) when we derive the DV EM scheme. Then, we combine the two extensions of EM we derived (i.e., OtO EM and DV EM) to obtain a comprehensive EM scheme (OtO+DV EM) that incorporates constraints on both sources and observed variables into the estimation framework.

A. Deriving the Likelihood

In order to derive a likelihood function that considers constraints in the form of dependencies between observed variables, we first divide the N observed variables in our social sensing model into G independent groups, where each independent group contains variables that are related by some local constraints (e.g., gas price of stations in the same neighborhood could be highly correlated). Consider group g, where there are k dependent variables $g_1, ..., g_k$. Let $p(z_{g_1}, ..., z_{g_k})$ represent the joint probability distribution of the k variables and let $\mathcal{Y}_g$ represent all possible combinations of values of $g_1, ..., g_k$. For example, when there are only two variables, $\mathcal{Y}_g = [(1,1), (1,0), (0,1), (0,0)]$. Note that, we assume that $p(z_{g_1}, ..., z_{g_k})$ is known or can be estimated from prior knowledge. The new likelihood function $L(\theta; X, Z)$ that considers the aforementioned constraints is:

$$L(\theta; X, Z) = \prod_{g \in G} p(X_g, Z_g|\theta) = \prod_{g \in G} p(Z_g) \times p(X_g|Z_g, \theta)$$
$$= \prod_{g \in G} \left\{ \sum_{g_1,...,g_k \in \mathcal{Y}_g} p(z_{g_1}, ..., z_{g_k}) \prod_{i \in M} \prod_{j \in c_g} \alpha_{i,j} \right\} \tag{10}$$

where $\alpha_{i,j}$ is the same as defined in Equation (9) and c_g represents the set of variables belonging to the independent group g. Compared to our previous effort [36], the new likelihood function is formulated with independent groups as units (instead of single independent variables). The joint probability distribution of all dependent variables within a group is used to replace the distribution of a single variable. This likelihood function is therefore more general, but reduces to the previous form in the special case where each group is composed of only one variable.

Using the above likelihood function, we can derive the corresponding E-Step and M-Step of DV EM and OtO+DV EM schemes. The detailed derivations are shown in Section IX-B.

B. The OtO+DV Algorithm

In summary, the OtO+DV EM scheme incorporates constraints on both sources and observed variables. The inputs to the algorithm are (i) the claim matrix SC, (ii) the knowledge matrix SK, and (iii) the *joint distribution* for each group of dependent variables, collectively represented by set JD. The output is the maximum likelihood estimate of source reliability and claim correctness. The OtO+DV EM pseudocode is shown in Algorithm 2.

V. Evaluation

In this section, we evaluate the performance of our new reliable social sensing schemes that incorporate "opportunity to observe" constraints on sources (OtO EM) and dependency constraints on observed variables (DV EM), as well as the comprehensive scheme (OtO+DV EM) that combines both. We compare their performance to the state of the art scheme from previous work [36] (regular EM) through a real world social sensing application. The purpose of the application is to map locations of traffic lights and stop signs on campus of the University of Illinois (in the city of Urbana-Champaign).

We use the dataset from a smartphone-based vehicular sensing testbed, called SmartRoad [16], where vehicle-resident Android smartphones record their GPS location traces as the cars are driven around by participants. The GPS readings include samples of the instantaneous latitude- longitude location, speed and bearing of the vehicle, with a sampling rate of 1 second. We aim to show that even very unreliable sensing of traffic lights and stop signs can result in a good final map once our algorithm is applied to these claims to determine their odds of correctness. Hence, an intentionally simple-minded application scenario was designed to identify stop signs and traffic lights from GPS data.

[1] In practice, we can run the algorithm until the difference of estimation parameter between consecutive iterations becomes insignificant.

Algorithm 2 Expectation Maximization Algorithm with Constraints on Both Sources and Observed Variables (OtO+DV EM)

Initialize θ with random values between 0 and 1
while $\theta^{(t)}$ does not converge **do**
  for $j = 1 : N$ **do**
    compute $Z(t, j)$ as the *marginal distribution* of the joint probability as shown in Equation (17)
  end for
  $\theta^{(t+1)} = \theta^{(t)}$
  for $i = 1 : M$ **do**
    compute $a_i^{(t+1)}, b_i^{(t+1)}, d_j^{(t+1)}$ based on Equation (18)
    update $a_i^{(t)}, b_i^{(t)}, d_j^{(t)}$ with $a_i^{(t+1)}, b_i^{(t+1)}, d_j^{(t+1)}$ in $\theta^{(t+1)}$
  end for
  $t = t + 1$
end while
Let Z_j^c = converged value of $Z(t, j)$
Let a_i^c = converged value of $a_i^{(t)}$; b_i^c = converged value of $b_i^{(t)}$; d_i^c = converged value of $d_j^{(t)}$ $\quad j \in \mathcal{C}_i$
for $j = 1 : N$ **do**
  if $Z_j^c \geq threshold$ **then**
    C_j is true
  else
    C_j is false
  end if
end for
for $i = 1 : M$ **do**
  calculate t_i^* from a_i^c, b_i^c and d_j^c
end for
Return the classification on variables and reliability estimation of sources

Specifically, in our experiment, if a vehicle waits at a location for 15-90 seconds, the application concludes that it is stopped at a traffic light and issues a traffic-light claim (i.e., a claim that a traffic light is present at that location and bearing). Similarly if it waits for 2-10 seconds, it concludes that it is at a stop sign and issues a stop-sign claim (i.e., a claim that a stop sign is present at that location and bearing). If the vehicle stops for less than 2 seconds, for 10-15 seconds, or for more than 90 seconds, no claim is made. Claims were reported by each source to a central data collection point.

Clearly the claims defined above are very error-prone due to the simple-minded nature of the "sensor" and the complexity of road conditions and driver's behaviors. Moreover, it is hard to quantify the reliability of sources without a training phase that compares measurements to ground truth. For example, a car can stop elsewhere on the road due to a traffic jam or crossing pedestrians, not necessarily at locations of traffic lights and stop signs. Also, a car does not stop at traffic lights that are green and a careless driver may pass stop signs without stopping. The question addressed in the evaluation is whether knowledge of constraints, as described in this paper, helps improve the accuracy of stop sign and traffic light estimation from such unreliable measurements in this case study.

Hence, we applied the different estimation approaches developed in this paper along with the constraints from the physical world on the noisy data to identify the correct locations of traffic lights and stop signs and compute the reliability of participants. One should note that location granularity here is of the order of half a city block. This ensures that stop sign and traffic light claims are attributed to the correct intersections. Most GPS devices easily attain such granularity. Therefore, the authors do not expect location errors to be of concern. For evaluation purposes, we manually collected the ground truth locations of traffic lights and stop signs.

In the experiment, 34 people (sources) were invited to participate and 1,048,572 GPS readings (around 300 hours of driving) were collected. A total of 4865 claims were generated by the phones, of which 3303 were for stop signs and 1562 were for traffic lights, collectively identifying 369 distinct locations. The elements S_iC_j of the claim matrix were set according to the claims extracted from each source vehicle.

We observed that traffic lights at an intersection are always present in all directions. Hence, when processing traffic light claims, we ignored vehicle bearing. However, stop signs at an intersection have a few possible scenarios. For example, (i) a stop sign may be present in each possible direction (e.g., All-Way stop); (ii) two stop signs may exist on one road whereas no stop sign exist on the other road (e.g., a main road intersecting with a small road); or (iii) two stop signs may exist for one road and one stop sign for the other road (e.g., a two-way road intersecting with a one way road). Hence, in claims regarding stop signs the bearing is important. We bin bearing into four main directions. A different Boolean variable is created for each direction.

A. Opportunity to Observe

In this subsection, we first evaluate the performance of the OtO EM scheme. For the OtO EM scheme, we used the recorded GPS traces of each vehicle to determine whether it actually went to a specific location or not (i.e., decide whether a source has an opportunity to observe a given variable or not).There are 54 actual traffic lights and 190 stop signs covered by the data traces collected.

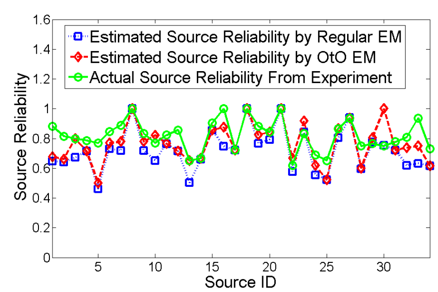
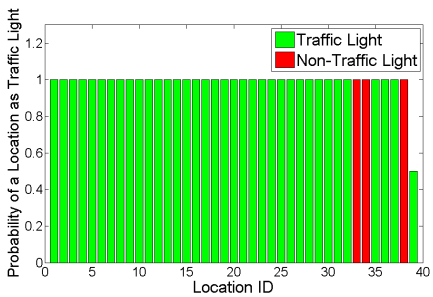


Fig. 1. Source Reliability Estimation of OtO EM in the Case of Traffic Lights

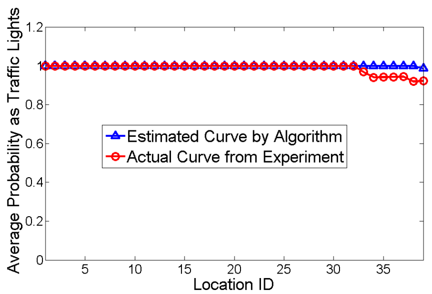
Figure 1 compares the source reliability estimated by both the OtO EM and regular EM schemes to the actual source reliability computed from ground truth. We observed that the OtO EM scheme stays closer to the actual results for most of the sources (i.e., OtO EM estimation error is smaller than regular EM for about 74% of sources).

Next, we explore the accuracy of identifying traffic lights by the new scheme. It may be tempting to confuse the problem with one of classification and plot ROC curves or confusion matrices. This would not be appropriate, however, because the output of our algorithm is not a classification label, but rather a probability that the labeled entity (e.g., a traffic light) exists at a given location. Some locations are associated with a higher probability than others. Hence, what is needed is an estimate of how well the computed probabilities match ground truth.

Accordingly, Figure 2 and Figure 3 show the accuracy of identifying traffic lights by the OtO EM scheme (Figure 2) and the regular EM scheme (Figure 3). The horizontal axis shows location IDs where lights were indicated by the respective schemes, sorted by their probability (as computed from the corresponding EM variant) from high to low. Hence, one should expect that lower-numbered locations be true positives, whereas high-numbered locations may contain increasingly more false positives as the scheme assigns a lower probability that those contain traffic lights.

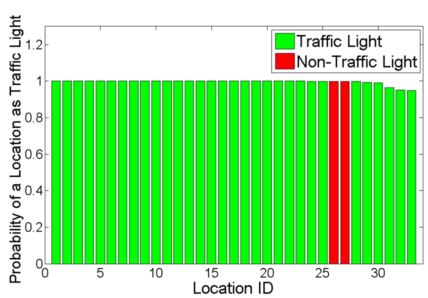


(a) Locations Identified as Traffic Lights

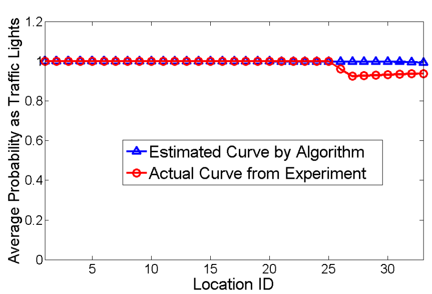


(b) Average Probability as Traffic Lights

Fig. 2. Claim Classification of OtO EM in the Case of Traffic Lights



(a) Locations Identified as Traffic Lights



(b) Average Probability as Traffic Lights

Fig. 3. Claim Classification of Regular EM in the Case of Traffic Lights

Figure 2(a) shows the actual status of each location. A green (light) bar is a true positive, whereas a red (dark) one is a false positive. As expected, we observe that most of the traffic light locations identified by the OtO EM scheme are true positives. False positives occur only later on at higher-numbered (i.e., lower probability) locations. Additionally, it is interesting to compare the probability of finding traffic lights at the indicated locations, as computed by our algorithm, to the probability computed empirically for the same locations by counting how many of them have actual lights. Figure 2(b) shows this comparison. Specifically, it compares the average probability to the empirical probability, computed for the first n locations to have traffic lights, where n is the location index on the horizontal axis. We observe that the estimation results of OtO EM follow quite well the empirical ones.

	Regular EM	OtO EM
Average Source Reliability Estimation Error	10.19%	7.74%
Number of Correctly Identified Traffic Lights	31	36
Number of Mis-Identified Traffic Lights	2	3

TABLE I. PERFORMANCE COMPARISON BETWEEN REGULAR EM VS OTO EM IN CASE OF TRAFFIC LIGHTS

Similarly, results for the regular EM scheme are reported in Figure 3. We observe that the OtO EM scheme is able to find five more traffic light locations compared to the regular EM scheme. The detailed comparison results between two schemes are given in Table I.

We repeated the above experiments for stop sign identification and observed that the OtO EM scheme achieves a more significant performance gain in both participant reliability estimation and stop sign classification accuracy compared to the regular EM scheme. The reason is: stop signs are scattered in town and the odds that a vehicle's path covers most of the stop signs are usually small. Hence, having the knowledge of whether a source had an opportunity to observe a variable is very helpful. However, we do find in general that the identification of stop signs is more challenging than that of traffic lights. There are several reasons for that. Namely, (i) the claims for stop signs are sparser because stops signs are typically located on smaller streets, so the chances of different cars visiting the same stop sign are lower than that for traffic lights, (ii) cars often stop briefly at non-stop sign locations, which our sensors mis-interpret for stop signs, and (iii) when cars want to make a turn after the stop sign, cars' bearings are often not well aligned with the directions of stop signs, which causes errors since stop-sign claims are bearing-sensitive.

Figure 4 compares source reliability computed by the OtO EM and regular EM schemes. The actual reliability is computed from experiment data similarly as we did for traffic lights. We observe that source reliability is better estimated by the OtO EM scheme compared to the regular EM scheme.

Figure 5 and Figure 6 show the true positives and false positives in recognizing stop signs. We observe the OtO EM scheme actually finds twelve more correct stop sign locations and reduces one false positive location compared to the regular EM scheme. The detailed comparison results are given in

	Regular EM	OtO EM	DV EM	DV+OtO EM
Average Source Reliability Estimation Error (Full Dataset)	25.34%	16.75%	15.99%	11.98%
Number of Correctly Identified Stop Signs (Full Dataset)	127	139	141	146
Number of Mis-Identified Stop Signs (Full Dataset)	25	24	29	25
Average Source Reliability Estimation Error (75% Dataset)	36.44%	18.2%	18.0%	15.29%
Number of Correctly Identified Stop Signs (75% Dataset)	92	101	111	116
Number of Mis-Identified Stop Signs (75% Dataset)	18	23	30	29

TABLE II. PERFORMANCE COMPARISON OF REGULAR EM, OTO EM, DV EM AND DV+OTO EM IN CASE OF STOP SIGNS

Table II. To further investigate the effects of data sparsity on different schemes, we repeat the above experiments using only 75% of the claims we collected. Results are also reported in Table II. Additionally, we observe that, for both EM schemes, the actual probability of finding stop signs at the indicated locations stays close to but slightly less than the estimated probability by our algorithms. The reasons of such deviation can be explained by the aftermentioned short wait behaviors at non-stop sign locations in real world scenarios.

B. Dependent Variables

In this subsection, we evaluated our extensions that consider dependency constraints (DV EM), and the comprehensive OtO+DV EM scheme. While the earlier discussion treated stop signs as independent variables, this is not strictly so. The existence of stop signs in different directions (bearings) is in fact quite correlated. We empirically computed those correlations for Urbana-Champaign and assumed that we knew them in advance. Clearly, the more "high-order" correlations are considered, the more information is given to improve performance of algorithm. To assess the effect of "minimal" information (which would be a "worst-case" improvement for our scheme), in this paper we consider pairwise correlations only. Hence, the joint distribution of co-existence of (two) stop signs in opposite directions at an intersection was computed. It is presented in Table III, and was used as input to the DV EM scheme.

Figure 7 shows the accuracy of source reliability estimation, when these constraints are used. We observe that both DV EM and DV+OtO EM scheme track the source reliability very well (the estimation error of the two EM schemes improved 9.4% and 13.4% repsectively compared to the regular EM scheme).

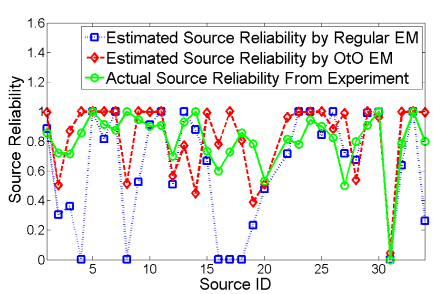
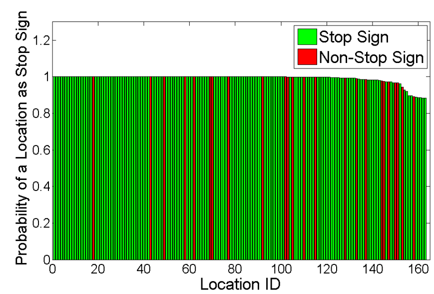
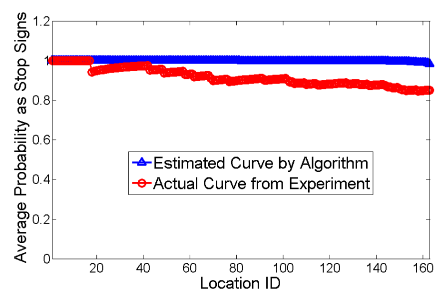


Fig. 4. Source Reliability Estimation of OtO EM in the Case of Stop Signs



(a) Locations Identified as Stop Signs



(b) Average Probability as Stop Signs

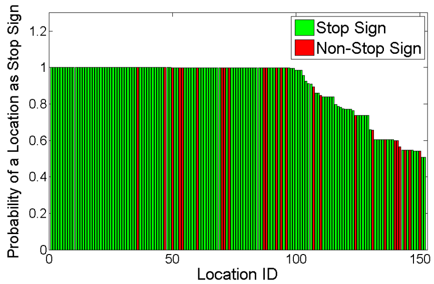
Fig. 5. Claim Classification of OtO EM in the Case of Stop Signs

A = stop sign 1 exists; B = stop sign 2 exists	Percentage
p(A,B)	36%
p(not A, not B)	49%
p(A,not B) = p(not A, B)	7.5%

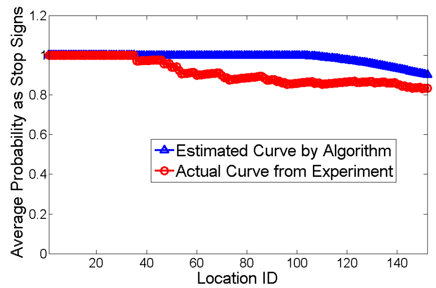
TABLE III. DISTRIBUTION OF STOP SIGNS IN OPPOSITE DIRECTIONS

The true positives and false positives for stop signs are shown in Figure 8 and Figure 9. Observe that the DV EM scheme finds 14 more correct stop sign locations. The DV+OtO EM scheme performed the best, it finds the most stop sign locations (i.e., 19 more than regular EM, 5 more than DV EM) while keeping the false positives the least (i.e., the same as regular EM and 4 less than DV EM). The detailed comparison results are given in Table II.

Additionally, we observe that, for the DV+OtO EM scheme, the estimated probability of finding stop signs is much closer to the empirically computed probability, compared to other EM schemes we discussed. This is because we explicitly considered both dependency constraints and the "opportunity



(a) Locations Identified as Stop Signs



(b) Average Probability as Stop Signs

Fig. 6. Claim Classification of Regular EM in the Case of Stop Signs

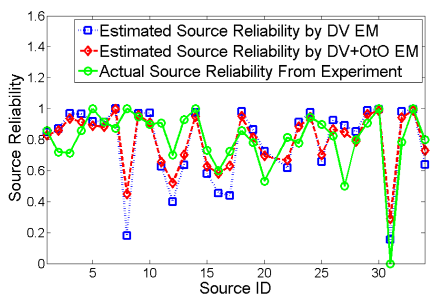
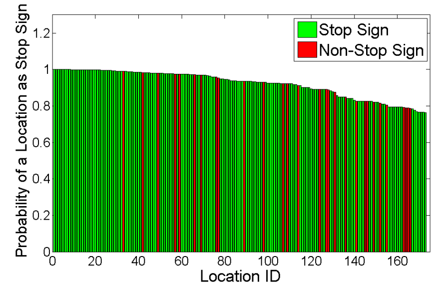
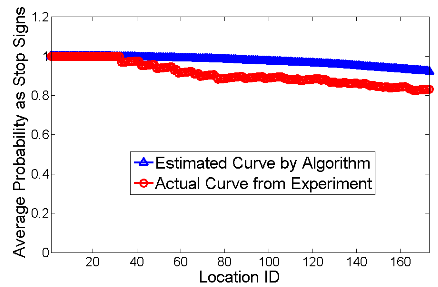


Fig. 7. Source Reliability Estimation of DV and DV+OtO EM in the Case of Stop Signs

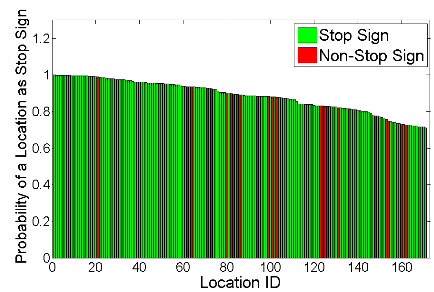


(a) Locations Identified as Stop Signs

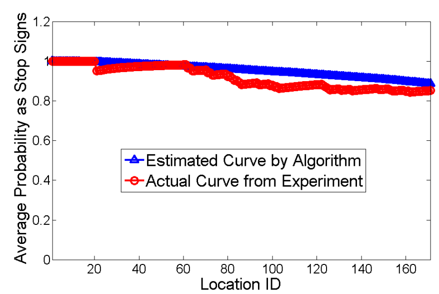


(b) Average Probability as Stop Signs

Fig. 8. Claim Classification of DV EM in the Case of Stop Signs



(a) Locations Identified as Stop Signs



(b) Average Probability as Stop Signs

Fig. 9. Claim Classification of DV+OtO EM in the Case of Stop Signs

to observe” for sources in the DV+OtO EM scheme.

VI. Discussion and Limitations

This paper presented a maximum likelihood estimation framework for exploiting the physical world constraints (i.e., source locations and observed variable dependencies) to improve the reliability of social sensing. Some limitations exist that offer directions for future work.

First, we did not explicitly model the time dimension of the problem in our current framework. This is mainly because our current application involves the detection of fixed infrastructure (e.g., stop signs and traffic lights). Time is less relevant in such context. Hence, opportunity to observe is only a function of source location, and observed variable dependencies are not likely to change over time. It would be interesting to consider time constraints in our future models. In systems where the state of the environment may change over time, when we consider the opportunity to observe, it is not enough for the source to have visited a location of interest. It is also important that the source visits that location within a certain time bound during which the state of the environment has not changed. Similarly, when we consider observed variable dependencies, it is crucial that dependencies of observed variables remain stable within a given time interval and that we have an efficient way to quickly update our estimation on such dependencies as time goes by.

Second, we assume sources will only report claims for the places they have been to (e.g., cars only generate stop sign claims on the streets their GPS traces covered). Hence, it makes sense to "penalize" sources for not making claims for some clearly observable variables based on their locations. However, many other factors might also influence the opportunity of users to generate claims in real-world social sensing applications. Some of these factors are out of user's control. For example, in some geo-tagging applications, participants use their phones to take photos of locations of interest. However, this approach might not work at some places due to "photo prohibited" signs or privacy concerns. Source reliability penalization based on visited locations might not be appropriate in such context. It is interesting to extend the notion of location-based opportunity-to-observe in our model to consider different types of source constrains in other social sensing applications.

Third, we do not assume "Byzantine" sources in our model (e.g., cars will not cheat in reporting the their GPS coordinates). However, in some crowd-sensing applications, sources can intentionally report incorrect locations (e.g., Google' Ingress). Different techniques have been developed to detect and address location cheating attacks on both mobile sensing applications [15] and social gaming systems [22]. These techniques can be used along with our schemes to solve the truth estimation problem in social sensing applications where source's reliability is closely related to their locations. Moreover, it is also interesting to further investigate the robustness of our scheme with respect to the percentage of cheating sources in the system.

Last, we assume that the joint probability distribution of dependent variables is known or can be estimated from prior knowledge. This might not be possible for all social sensing applications. Clearly, the approach in the current paper would not apply if nothing was known about spatial correlations in environmental state. Additionally, the scale of current experiment is relatively small. We are working on new social sensing applications, where we can test our models at a larger scale.

VII. Related Work

Social sensing emerged recently as a key area of research in sensor networks due to the great increase in the number of mobile sensors owned by individuals (e.g., smart phones), the proliferation of Internet connectivity, and the fast growth in mass dissemination media (e.g., Twitter, Facebook, and Flickr, to name a few). Social sensing applications can be seen as a broad set of applications, where humans play a key role in the data collection system by acting as sensor carriers [21] (e.g., opportunistic sensing), sensor operators [4] (e.g., participatory sensing) or sensor themselves. An early overview of social sensing applications is described in [1]. Examples of early systems include CenWits [17], CarTel [18], BikeNet [8], and CabSense [32]. Recent work explored privacy [27], energy-efficient context sensing [25], and social interaction aspects [30].

In this paper, we are particularly interested in the data reliability aspect of social sensing. The social sensing paradigm draws strength from its inclusive nature; anyone can particulate in sensing and the barrier to entry is low. Such openness is a coin of two sides: on one hand, it greatly increases the availability of information and the diversity of sources. On the other hand, it introduces the problem of understanding the reliability of the contributing sources and ensuring the quality of the information collected. Solutions such as the Trusted Platform Module (TPM) [11] and YouProve [12] can be used to provide a certain level of assurance that the source device is running authentic software. However, this is not sufficient in sensing applications because it does not guarantee the integrity of use, physical context, and environment. For example, a smart phone application intended to measure vehicular traffic speed can be turned on when the user is on a bicycle, hence resulting in unrepresentative measurements.

There exists a good amount of work in the data mining and machine learning communities on the topic of fact-finding, which addresses the challenge of assertaining correctness of data from unreliable sources. For example, Hubs and Authorities [19] presents a simple empirical model that jointly computes the credibility of information sources and their claims in an iterative fashion. Similar work includes the TruthFinder [42] and the Investment, PooledInvestment and Average·Log algorithms [26]. Additional frameworks have been proposed to enhance this basic model to consider the dependency between sources [7], a source's varying expertise across different topics, and the notion of hardness of facts asserted by sources [9]. Pasternack et al. [26] further proposed a comprehensive framework to incorporate the prior knowledge concerning the claims (in the form of first-order logic) into fact-finding to leverage what the user already knows.

More recent work came up with new fact-finding algorithms by applying techniques in statistics and estimation theory to do trust analysis of information network in a principled way. Zhao et al. [43] presented Bayesian network model to handle different types of errors made by sources and merge multi-valued attribute types of entities in data integration systems. Wang et al. [36], [37] proposed a maximum likelihood estimation framework that offers a joint estimation on source reliability and claim correctness based on a set of general simplifying assumptions. In their following work, Wang et al. further quantified the accuracy of the maximum likelihood estimation (MLE) [35], [38] and extended their framework to handle streaming data [34]. The approach was compared to several state-of-the-art previous fact-finders and was shown to outperform them in estimation accuracy [36]. Accordingly, we only compare our new extensions to the winning approach from prior art.

Finally, physical constraints and models (both spatial and temporal) have been extensively studied in the wireless sensor network (WSN) community. They have often been used to reduce resource consumption by leveraging knowledge of the physical model or constraint to reduce data transmission needs. Compression and coding schemes were proposed to reduce the data redundancy in the space domain [31], [41]. Temporal correlations were exploited to reduce network load while offering compression quality guarantees [2], [13]. The contribution of our work lies in incorporating the constraints from the physical world into a framework for *improving estimation accuracy* as opposed to *reducing resource cost*. The underlying insight is the same: knowledge of physical correlations and constraints between variables reduces problem dimensionality. Prior WSN work harvests such reduction to correspondingly reduce data transmission needs. In contrast, we harvest it to improve noise elimination at the same resource cost.

VIII. Conclusion

This paper presented a framework for incorporating source and claim constraints that arise from physical knowledge (of source locations and observed variable dependencies) into maximum-likelihood analysis to improve the accuracy of social sensing. The problem addressed was one of jointly assessing the probability of correctness of claims and the reliability of their sources by exploiting physical constraints and data provenance relations to better estimate the likelihood of reported observations. An expectation maximization scheme was described that arrives at a maximum likelihood solution. The performance of the new algorithm was evaluated through a real world social sensing application. Results show a significant reduction in estimation error of both source reliability and claim correctness thanks to the exploitation of physical constraints.

IX. Appendix

A. Derivation of the E-step and M-step of OtO EM

Having formulated the new likelihood function to account for the source constraints in the previous subsection, we can now plug it into the Q function defined in Equation (6) of Expectation Maximization. The E-step can be derived as follows:

$$\begin{aligned} Q\left(\theta|\theta^{(t)}\right) &= E_{Z|X,\theta^{(t)}}[\log L(\theta;X,Z)] \\ &= \sum_{j=1}^{N}\Bigg\{p(z_j=1|X_j,\theta^{(t)}) \times \sum_{i\in\mathcal{S}_j}(\log\alpha_{i,j}+\log d_j) \\ &+ p(z_j=0|X_j,\theta^{(t)}) \times \sum_{i\in\mathcal{S}_j}(\log\alpha_{i,j}+\log(1-d_j))\Bigg\} \end{aligned} \tag{11}$$

where $p(z_j=1|X_j,\theta^{(t)})$ represents the conditional probability of the variable C_j to be true given the claim matrix related to the j^{th} claim and current estimate of θ. We represent $p(z_j = 1|X_j,\theta^{(t)})$ by $Z(t,j)$ since it is only a function of t and j. $Z(t,j)$ can be further computed as:

$$\begin{aligned} Z(t,j) &= p(z_j=1|X_j,\theta^{(t)}) \\ &= \frac{p(z_j=1;X_j,\theta^{(t)})}{p(X_j,\theta^{(t)})} \\ &= \frac{p(X_j,\theta^{(t)}|z_j=1)p(z_j=1)}{p(X_j,\theta^{(t)}|z_j=1)p(z_j=1)+p(X_j,\theta^{(t)}|z_j=0)p(z_j=0)} \\ &= \frac{\prod_{i\in\mathcal{S}_j}\alpha_{i,j}\times d_j^{(t)}}{\prod_{i\in\mathcal{S}_j}\alpha_{i,j}\times d_j^{(t)}+\prod_{i\in\mathcal{S}_j}\alpha_{i,j}\times(1-d_j^{(t)})} \end{aligned} \tag{12}$$

Note that, in the E-step, we continue to only consider sources who observe a given variable while computing the likelihood of reports regarding that variable.

In the M-step, we set the derivatives $\frac{\partial Q}{\partial a_i} = 0$, $\frac{\partial Q}{\partial b_i} = 0$, $\frac{\partial Q}{\partial d_j} = 0$. This gives us the θ^* (i.e., $a_1^*, a_2^*, ...a_M^*; b_1^*, b_2^*, ...b_M^*; d_1^*, d_2^*, ..., d_N^*$) that maximizes the $Q\left(\theta|\theta^{(t)}\right)$ function in each iteration and is used as the $\theta^{(t+1)}$ of the next iteration.

$$\begin{aligned} a_i^{(t+1)} &= a_i^* = \frac{\sum_{j\in SJ_i}Z(t,j)}{\sum_{j\in\mathcal{C}_i}Z(t,j)} \\ b_i^{(t+1)} &= b_i^* = \frac{\sum_{j\in SJ_i}(1-Z(t,j))}{\sum_{j\in\mathcal{C}_i}(1-Z(t,j))} \\ d_j^{t+1} &= d_j^* = Z(t,j) \\ d_i^* &= \frac{\sum_{j\in\mathcal{C}_i}Z(t,j)}{|\mathcal{C}_i|} \end{aligned} \tag{13}$$

where $\mathcal{C}_i$ is set of variables source S_i observes according to the knowledge matrix SK and $Z(t,j)$ is defined in Equation (12). SJ_i is the set of variables the source S_i actually claims in the claim matrix SC. We note that, in the computation of a_i and b_i, the silence of source S_i regarding some variable C_j is interpreted differently depending on whether S_i *observed* it or not. This reflects that the opportunity to observe has been incorporated into the M-Step when the estimation parameters of sources are computed. The resulting OtO EM algorithm is summarized in the subsection below.

B. Derivation of E-Step and M-Step of DV and OtO+DV EM

Given the new likelihood function of the DV EM scheme defined in Equation (10), the E-step becomes:

$$\begin{aligned} Q\left(\theta|\theta^{(t)}\right) &= E_{Z|X,\theta^{(t)}}[\log L(\theta;X,Z)] \\ &= \sum_{g\in G}p(z_{g_1},...,z_{g_k}|X_g,\theta^{(t)}) \\ &\times\left\{\sum_{i\in M}\sum_{j\in c_g}\log\alpha_{i,j}+\log p(z_{g_1},...,z_{g_k})\right\} \end{aligned} \tag{14}$$

where $p(z_{g_1},...,z_{g_k}|X_g,\theta^{(t)})$ represents the conditional joint probability of all variables in independent group g (i.e., $g_1,...,g_k$) given the observed data regarding these variables and the current estimation of the parameters. $p(z_{g_1},...,z_{g_k}|X_g,\theta^{(t)})$ can be further computed as follows:

$$\begin{aligned} &p(z_{g_1},...,z_{g_k}|X_g,\theta^{(t)}) \\ &= \frac{p(z_{g_1},...,z_{g_k};X_g,\theta^{(t)})}{p(X_g,\theta^{(t)})} \\ &= \frac{p(X_g,\theta^{(t)}|z_{g_1},...,z_{g_k})p(z_{g_1},...,z_{g_k})}{\sum_{g_1,...,g_k\in\mathcal{Y}_g}p(X_g,\theta^{(t)}|z_{g_1},...,z_{g_k})p(z_{g_1},...,z_{g_k})} \\ &= \frac{\prod_{i\in M}\prod_{j\in c_g}\alpha_{i,j}p(z_{g_1},...,z_{g_k})}{\sum_{g_1,...,g_k\in\mathcal{Y}_g}\prod_{i\in M}\prod_{j\in c_g}\alpha_{i,j}p(z_{g_1},...,z_{g_k})} \end{aligned} \tag{15}$$

We note that $p(z_j = 1|X_j,\theta^{(t)})$ (i.e., $Z(t,j)$), defined as the probability that C_j is true given the observed data and the current estimation parameters, can be computed as the *marginal distribution* of the joint probability of all variables in the independent claim group g that variable C_j belongs to (i.e., $p(z_{g_1},...,z_{g_k}|X_g,\theta^{(t)}) \quad j\in c_g$). We also note that, for the worst case where N variables fall into one independent group, the computational load to compute this marginal grows exponentially with respect to N. However, as long as the constraints on observed variables are localized, our approach stays scalable, independently of the total number of estimated variables.

In the M-step, as before, we choose θ^* that maximizes the $Q\left(\theta|\theta^{(t)}\right)$ function in each iteration to be the $\theta^{(t+1)}$ of the next iteration. Hence:

$$\begin{aligned} a_i^{(t+1)} &= a_i^* = \frac{\sum_{j \in SJ_i} Z(t,j)}{\sum_{j=1}^{N} Z(t,j)} \\ b_i^{(t+1)} &= b_i^* = \frac{\sum_{j \in SJ_i} (1 - Z(t,j))}{\sum_{j=1}^{N} (1 - Z(t,j))} \\ d_j^{t+1} &= d_j^* = Z(t,j) \end{aligned} \tag{16}$$

where $Z(t,j) = p(z_j = 1|X_j, \theta^{(t)})$. We note that for the estimation parameters, a_i and b_i, we obtain the same expression as for the case of independent variables. The reason is that sources report variables independently of the form of constraints between these variables.

Next, we combine the two EM extensions (i.e., OtO EM and DV EM) derived so far to obtain a comprehensive EM scheme (OtO+DV EM) that considers constraints on both sources and observed variables. The corresponding E-Step and M-Step are shown below:

$$\begin{aligned} & p(z_{g_1},...,z_{g_k}|X_g,\theta^{(t)}) = \frac{p(z_{g_1},...,z_{g_k};X_g,\theta^{(t)})}{p(X_g,\theta^{(t)})} \\ & = \frac{p(X_g,\theta^{(t)}|z_{g_1},...,z_{g_k})p(z_{g_1},...,z_{g_k})}{\sum_{g_1,...,g_k \in \mathcal{Y}_g} p(X_g,\theta^{(t)}|z_{g_1},...,z_{g_k})p(z_{g_1},...,z_{g_k})} \\ & = \frac{\prod_{i \in \mathcal{S}_j} \prod_{j \in c_g} \alpha_{i,j} p(z_{g_1},...,z_{g_k})}{\sum_{g_1,...,g_k \in \mathcal{Y}_g} \prod_{i \in \mathcal{S}_j} \prod_{j \in c_g} \alpha_{i,j} p(z_{g_1},...,z_{g_k})} \\ & \text{where } \mathcal{S}_j : \text{Set of sources observes } C_j \end{aligned} \tag{17}$$

$$\begin{aligned} a_i^{(t+1)} &= a_i^* = \frac{\sum_{j \in SJ_i} Z(t,j)}{\sum_{j \in \mathcal{C}_i} Z(t,j)} \\ b_i^{(t+1)} &= b_i^* = \frac{\sum_{j \in SJ_i} (1 - Z(t,j))}{\sum_{j \in \mathcal{C}_i} 1 - Z(t,j))} \\ d_j^{t+1} &= d_j^* = Z(t,j) \\ & \text{where } \mathcal{C}_i \text{ is set of variables source } S_i \text{ observes} \end{aligned} \tag{18}$$

Acknowledgements

Research reported in this paper was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement W911NF-09-2-0053, DTRA grant HDTRA1-10-1-0120, and NSF grants CNS 09-05014 and CNS 10-35736. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

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

[2] A. Ali, A. Khelil, P. Szczytowski, and N. Suri. An adaptive and composite spatio-temporal data compression approach for wireless sensor networks. In *Proceedings of the 14th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems*, MSWiM '11, pages 67–76, New York, NY, USA, 2011. ACM.

[3] S. Barbarossa, G. Scutari, and S. Member. Decentralized maximum likelihood estimation for sensor networks composed of nonlinearly coupled dynamical systems. *IEEE Trans. Signal Process*, 55:3456–3470, 2007.

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

[5] M. Chang, L. Ratinov, and D. Roth. Structured learning with constrained conditional models. *Machine Learning*, 88(3):399–431, 6 2012.

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

[10] R. K. Ganti, N. Pham, H. Ahmadi, S. Nangia, and T. F. Abdelzaher. Greengps: a participatory sensing fuel-efficient maps application. In *MobiSys '10: Proceedings of the 8th international conference on Mobile systems, applications, and services*, pages 151–164, New York, NY, USA, 2010. ACM.

[11] P. Gilbert, L. P. Cox, J. Jung, and D. Wetherall. Toward trustworthy mobile sensing. In *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*, HotMobile '10, pages 31–36, New York, NY, USA, 2010. ACM.

[12] P. Gilbert, J. Jung, K. Lee, H. Qin, D. Sharkey, A. Sheth, and L. P. Cox. Youprove: authenticity and fidelity in mobile sensing. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, SenSys '11, pages 176–189, New York, NY, USA, 2011. ACM.

[13] A. Guitton, A. Skordylis, and N. Trigoni. Utilizing correlations to compress time-series in traffic monitoring sensor networks. In *Wireless Communications and Networking Conference, 2007.WCNC 2007. IEEE*, pages 2479–2483, 2007.

[14] L. He and R. Greenshields Ian. A nonlocal maximum likelihood estimation method for rician noise reduction in mr images. *Medical Imaging, IEEE Transactions on*, 28(2):165–172, 2009.

[15] W. He, X. Liu, and M. Ren. Location cheating: A security challenge to location-based social network services. In *Distributed Computing Systems (ICDCS), 2011 31st International Conference on*, pages 740–749, 2011.

[16] S. Hu, H. Liu, L. Su, H. Wang, and T. Abdelzaher. SmartRoad: A Mobile Phone Based Crowd-Sourced Road Sensing System. Technical report, University of Illinois at Urbana-Champaign, 08 2013. https://www.ideals.illinois.edu/handle/2142/45699.

[17] 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.

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

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

[20] E. Koukoumidis, L.-S. Peh, and M. Martonosi. Demo: Signalguru: leveraging mobile phones for collaborative traffic signal schedule advisory. In *Proceedings of the 9th international conference on Mobile systems, applications, and services*, MobiSys '11, pages 353–354, New York, NY, USA, 2011. ACM.

[21] N. D. Lane, E. Miluzzo, S. B. Eisenman, M. Musolesi, and A. T. Campbell. Urban sensing systems: Opportunistic or participatory, 2008.

[22] J. Martn de Valmaseda, G. Ionescu, and M. Deriaz. Trustpos model: Trusting in mobile users location. In F. Daniel, G. Papadopoulos, and P. Thiran, editors, *Mobile Web Information Systems*, volume 8093 of *Lecture Notes in Computer Science*, pages 79–89. Springer Berlin Heidelberg, 2013.

[23] E. Monte-Moreno, M. Chetouani, M. Faundez-Zanuy, and J. Sole-Casals. Maximum likelihood linear programming data fusion for speaker recognition. *Speech Commun.*, 51(9):820–830, Sept. 2009.

[24] M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, and P. Boda. Peir, the personal environmental impact report, as a platform for participatory sensing systems research. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, MobiSys '09, pages 55–68, New York, NY, USA, 2009. ACM.

[25] 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.

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

[27] N. Pham, R. K. Ganti, Y. S. Uddin, S. Nath, and T. Abdelzaher. Privacy-preserving reconstruction of multidimensional data maps in vehicular participatory sensing, 2010.

[28] C. A. Pope, M. Ezzati, and D. W. Dockery. Fine-particulate air pollution and life expectancy in the united states. *New England Journal of Medicine*, 360(4):376–386, 2009. PMID: 19164188.

[29] T. Proietti and L. Alessandra. Maximum likelihood estimation of time series models: the kalman filter and beyond. Mpra paper, University Library of Munich, Germany, 2012.

[30] 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.

[31] A. Scaglione and S. D. Servetto. On the interdependence of routing and data compression in multi-hop sensor networks. In *Proceedings of the 8th annual international conference on Mobile computing and networking*, MobiCom '02, pages 140–147, New York, NY, USA, 2002. ACM.

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

[33] M. Y. S. Uddin, H. Wang, F. Saremi, G.-J. Qi, T. Abdelzaher, and T. Huang. Photonet: A similarity-aware picture delivery service for situation awareness. In *Proceedings of the 2011 IEEE 32nd Real-Time Systems Symposium*, RTSS '11, pages 317–326, Washington, DC, USA, 2011. IEEE Computer Society.

[34] 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.

[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] D. Wang, L. M. Kaplan, and T. F. Abdelzaher. Maximum likelihood analysis of conflicting observations in social sensing. *ACM Transaction on Sensor Networks*, to appear.

[38] D. Wang, L. M. Kaplan, T. F. Abdelzaher, and C. C. Aggarwal. On credibility estimation tradeoffs in assured social sensing. *IEEE Journal on Selected Areas in Communications*, 31(6):1026–1037, 2013.

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

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

[41] Y. Xu and W. chien Lee. Exploring spatial correlation for link quality estimation in wireless sensor networks. In *in Proc. IEEE PerCom*, pages 200–211, 2006.

[42] 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.

[43] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *Proc. VLDB Endow.*, 5(6):550–561, Feb. 2012.

[44] P. Zhou, Y. Zheng, and M. Li. How long to wait?: predicting bus arrival time with mobile phone based participatory sensing. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*, MobiSys '12, pages 379–392, New York, NY, USA, 2012. ACM.