Real-Time Syst
DOI 10.1007/s11241-015-9238-8

# Reliable social sensing with physical constraints: analytic bounds and performance evaluation

**Dong Wang[1] · Tarek Abdelzaher[2] · Lance Kaplan[3] · Raghu Ganti[4] · Shaohan Hu[2] · Hengchang Liu[2,5]**



**Abstract** Correctness guarantees are at the core of cyber-physical computing research. While prior research addressed correctness of timing behavior and correctness of program logic, this paper tackles the emerging topic of assessing correctness of input data. This topic is motivated by the desire to crowd-source sensing tasks, an act we henceforth call *social sensing*, in applications with humans in the loop. A key challenge in social sensing is that the reliability of sources is generally unknown, which makes it difficult to assess the correctness of collected observations. To address this challenge, we adopt a cyber-physical approach, where assessment of correctness of individual observations is aided by knowledge of physical constraints on 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 reported what) and physical constraints. We also derive new analytic bounds that allow the social sensing applications to accurately quantify the estimation error of source reliability for given confidence levels.

✉ Dong Wang
dwang5@nd.edu

1 Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA

2 Department of Computer Science, University of Illinois at Urbana Champaign, Urbana, IL 61801, USA

3 Networked Sensing and Fusion Branch, US Army Research Labs, Adelphi, MD 20783, USA

4 IBM Research, Yorktown Heights, NY 10598, USA

5 Department of Computer Science, University of Science and Technology of China, Hefei 230027, Anhui, People's Republic of China

Published online: 20 August 2015

We evaluate the framework through both a real-world social sensing application and extensive simulation studies. The results demonstrate significant performance gains in estimation accuracy of the new algorithms and verify the correctness of the analytic bounds we derived.

**Keywords** Social sensing · Cyber-physical computing · Maximum likelihood estimation · Physical constraint · Analytic bounds

## 1 Introduction

Attainment of correctness guarantees lies at the core of cyber-physical computing research. Prior research focused on guarantees of *timing correctness* and guarantees of *functional correctness*. In contrast, this paper investigates guarantees on *data correctness*.

The paper is motivated by the proliferation of cyber-physical applications with humans in the loop. For example, humans are the drivers in transportation systems, the consumers in smart grid applications, the first responders in disaster response systems, and the decision makers for sustainable ecosystems. As such, they can play a pivotal role in monitoring and reporting system state; an act we call *social sensing*.

We refer by *social sensing* applications 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 (Wang 2015). The proliferation of mobile devices with sensors, such as smartphones, has significantly increased the popularity of social sensing. Recent applications include optimization of daily commute (Zhou et al. 2012), reduction of carbon footprint (Koukoumidis et al. 2011), disaster response (Huang et al. 2005; Wang and Huang 2015) and pollution monitoring (Mun et al. 2009), 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 (Yin et al. 2008; Pasternack and Roth 2010; Qi et al. 2013). However, none of these work considered the physical constraints in their solutions due to the lack of explicit physical components in their application scenarios. Considering the tight integration of human, cyber and physical components in cyber-physical systems, this paper describes algorithms and analytic bounds to improve the reliability of social sensing applications by exploiting *physical constraints*.

Following the methodology reported in our earlier conference publication (Wang et al. 2013b), we adopt a cyber-physical approach to the problem of assessing correctness of collected data and obtaining the analytic bounds on source reliability, 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 use a maximum likelihood estimation (MLE) framework to jointly compute both the reliability

of sources and the correctness of the data they report, such that the likelihood function is maximized. This framework was first reported in (Wang et al. 2012b), but without taking physical constraints into account. The advantage of maximum-likelihood estimation lies in the feasibility of computing rigorous estimation accuracy bounds (Wang et al. 2012a), hence not only arriving at the top hypothesis, but also quantifying how good it is.

In contrast to our prior work (Wang et al. 2012b, 2013b, 2014), in this extended journal version, we derive new analytic bounds on estimation error that allow estimating confidence intervals in the (originally unknown) source reliability values. To the best of our knowledge, the derived analytic bounds in this paper are the first ones that explicitly consider the physical constraints in social sensing applications. We show that the maximum likelihood estimate obtained is a lot more accurate than one that does not take physical constraints into account and the analytic bounds we obtained correctly quantify the errors in the maximum likelihood estimation.

Much prior research in cyber-physical systems (CPS) (Hunter et al. 2012; Tang et al. 2012) and estimation theory (He and Greenshields Ian 2009; Proietti and Alessandra 2012) considered filtering observations of continuous variables in a maximum-likelihood fashion to separate signal from noise. While continuous variables are common in cyber-physical computing, an important subset of CPS 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. In this paper, we focused on a discrete variable scenario and formulated a reliable social sensing problem.

Our work is related to machine learning literature on constrained conditional models (Pasternack and Roth 2010; Chang et al. 2012). Unlike that literature, we do not limit our approach to simple linear models (Chang et al. 2012) nor require that constraints and constraints be deterministic (Pasternack and Roth 2010). 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 Sects. 3 and 4), and (ii) incorporate *probabilistic* constraints.

Finally, contrary to work that focuses on maximum-likelihood estimation of continuous variables given continuous models of physical phenomena, which appears in both cyber-physical systems and data fusion literature (Hunter et al. 2012; Tang et al. 2012; Monte-Moreno et al. 2009), 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 both a real-world social sensing application and extensive simulation studies. Results show significant performance improvements in both source reliability and variable classification as well as the effectiveness of the

analytic bounds we derived, achieved by incorporating physical information into the estimation framework.

The rest of the paper is organized as follows. Section 2 formulates the problem of reliable social sensing. Sections 3 and 4 solve the problem while leveraging source constraints and observed variable constraints, respectively. The new analytic bounds to quantify the estimation errors in source reliability are shown in Sect. 5. Evaluation results are presented in Sect. 6. The discussion is presented in Sect. 7. We review the related work in Sect. 8. Finally, we conclude the paper in Sect. 9.

## 2 The problem formulation

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 whose value 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, $S_1, \ldots, S_M$, observe a set of $N$ binary variables, $C_1, \ldots, C_N$. The value of a variable $C_j$ can be either true or false. The true value represents the positive state of the variable while the false value represents the negative state. Each variable $C_j$ is also associated with a location, $L_j$. 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.

In this paper, we exploit 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 report 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 *reported observation* $S_iC_j$ if the source reported that the value of $C_j$ was true. We generically denote by $p(C_j = 1|x)$ and $p(C_j = 0|x)$ the conditional probability that the value of variable $C_j$ is indeed true or false, given $x$, respectively. We denote by $t_i$ the (unknown) probability that the value of a randomly chosen variable is true given that source $S_i$ reported it (to be true). Formally, $t_i$ is given by:

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

Note that $C_j$ in the definition of $t_i$ is an arbitrary variable so $t_i$ represents the probability that the value of a variable is true conditioned on the knowledge that source i has espoused the truthfulness of the variable. Hence, $t_i$ does not depend on the variable index $j$.

Different sources may report different numbers of observations. The probability that source $S_i$ reports an observation 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 an observation given that the value of 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 an observation when the value of the underlying variable is in reality false and the source observed it. More formally:

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

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

$$\begin{aligned} a_i &= p\left(S_iC_j|C_j = 1, S_i \text{ observes } C_j\right) \\ &= \frac{p\left(C_j = 1|S_iC_j, S_i \text{ observes } C_j\right) \times p\left(S_iC_j|S_i \text{ observes } C_j\right)}{p\left(C_j = 1|S_i \text{ observes } C_j\right)} \\ b_i &= p\left(S_iC_j|C_j = 0, S_i \text{ observes } C_j\right) \\ &= \frac{p\left(C_j = 0|S_iC_j, S_i \text{ observes } C_j\right) \times p\left(S_iC_j|S_i \text{ observes } C_j\right)}{p\left(C_j = 0|S_i \text{ observes } C_j\right)} \end{aligned} \tag{3}$$

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 $C_j$ in the definition of $d_i$ represents a variable randomly chosen from all variables observed by $S_i$. So $d_i$ does not depend on the variable index $j$. This probability describes the proportion of variables that source $S_i$ observes that happen to have true values. Note that, the probability that a source reports an observation is proportional to the number of variables reported by the source over the total number

of variables observed by the source. In this paper, we assume sources only report variables it has an opportunity to observe (e.g., a car will only report the observation of a traffic light location when the car has an opportunity to visit that location). Under this assumption, $t_i = p(C_j = 1|S_i C_j, S_i \text{ observes } C_j)$. Plugging these terms into the definition of $a_i$ and $b_i$, given in Eq. (3), we get the relationship between the terms we defined above:

$$a_i = \frac{t_i \times s_i}{d_i} \qquad b_i = \frac{(1 - t_i) \times s_i}{1 - d_i}$$
$$d_i = p\left(C_j = 1|S_i \text{ observes } C_j\right) \tag{4}$$

The input to our algorithm is: (i) the *observation matrix* $SC$, where $SC_{ij} = 1$ when source $S_i$ reports that the value of $C_j$ is true, and $SC_{ij} = 0$ otherwise; and (ii) the source's opportunities to observe represented by a *knowledge matrix* $SK$, where $SK_{ij} = 1$ when source $S_i$ has the opportunity to observe $C_j$ and $SK_{ij} = 0$ otherwise. The output of the algorithm is the probability that the value of variable $C_j$ is true, for each $j$ and the reliability $t_i$ of source $S_i$, for each $i$. More formally:

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

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:

$$\forall j, 1 \leq j \leq N : p(C_j = 1|SC, SK, JD)$$
$$\forall i, 1 \leq i \leq M : p(C_j = 1|S_i C_j) \tag{6}$$

Below, we solve the aforementioned problems using the expectation maximization (EM) algorithm. EM (Dempster et al. 1977) 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 $\theta$ 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 (Hogg et al. 2005):

- E-step: Compute the log likelihood function for the M-step

$$Q\left(\theta|\theta^{(n)}\right) = E_{Z|X,\theta^{(n)}}[\log L(\theta; X, Z)] \tag{7}$$

- M-step: Maximize the Q function in the E-step

$$\theta^{(n+1)} = \underset{\theta}{\text{argmax}}\ Q\left(\theta|\theta^{(n)}\right) \tag{8}$$

Following the approach described in our previous work (Wang et al. 2012b), we define a latent variable $z_j$ to denote our estimated value of variable $C_j$, for each $j$ (indicating whether the value of $C_j$ 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 observation matrix $SC$, where $X_j$ represents the $j$th column of the $SC$ matrix (i.e., reported observations of the $j$th variable by all sources). The parameter vector we want to estimate is $\theta = (a_1, a_2, \ldots, a_M; b_1, b_2, \ldots, b_M; d_1, d_2, \ldots, d_N)$.

## 3 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).

### 3.1 Deriving the likelihood

When we consider source constraints in the likelihood function, we assume sources only report variables they observe, and hence the probability of a source reporting 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 Sect. 4. Under these assumptions, the new likelihood function $L(\theta; X, Z)$ that incorporates the source constraints is given by:

$$\begin{aligned} L(\theta; X, Z) &= p(X, Z|\theta) \\ &= \prod_{j=1}^{N} p(z_j) \times p(X_j|z_j, \theta) \\ &= \prod_{j=1}^{N} \prod_{i \in \mathcal{S}_j} p(z_j) \times \alpha_{i,j} \\ &\quad \text{where } \mathcal{S}_j : \text{Set of sources observed } C_j \end{aligned} \tag{9}$$

where

$$p(z_j) = \begin{cases} d_j & z_j = 1 \\ (1 - d_j) & z_j = 0 \end{cases}$$

$$\alpha_{i,j} = \begin{cases} a_i & z_j = 1,\ S_i C_j = 1 \\ (1 - a_i) & z_j = 1,\ S_i C_j = 0 \\ b_i & z_j = 0,\ S_i C_j = 1 \\ (1 - b_i) & z_j = 0,\ S_i C_j = 0 \end{cases} \tag{10}$$

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 (Wang et al. 2012b). This change allows us to

nicely incorporate the source constraints (name, source opportunity to observe) into the MLE 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 Sect. 1.

### 3.2 The OtO EM algorithm

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

Initialize $\theta$ with random values between 0 and 1
**while** $\theta^{(n)}$ does not converge **do**
  **for** $j = 1 : N$ **do**
    compute $Z(t, j)$ based on Eq. (23)
  **end for**
  $\theta^{(n+1)} = \theta^{(n)}$
  **for** $i = 1 : M$ **do**
    compute $a_i^{(n+1)}, b_i^{(n+1)}, d_j^{(n+1)}$ based on Eq. (24)
    update $a_i^{(n)}, b_i^{(n)}, d_j^{(n)}$ with $a_i^{(n+1)}, b_i^{(n+1)}, d_j^{(n+1)}$ in $\theta^{(n+1)}$
  **end for**
  $t = t + 1$
**end while**
Let $Z_j^c$ = converged value of $Z(t, j)$
Let $a_i^c$ = converged value of $a_i^{(n)}$; $b_i^c$ = converged value of $b_i^{(n)}$; $d_i^c$ = converged value of $d_j^{(n)}$ $j \in \mathcal{C}_i$
**for** $j = 1 : N$ **do**
  **if** $Z_j^c \geq threshold$ **then**
    the value of $C_j$ 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

In summary, the inputs to the OtO EM algorithm are (i) the observation 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 binary variable classification. Compared to the regular EM algorithm we derived in our previous work (Wang et al. 2012b), 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 $\theta$ with random values between 0 and 1. The algorithm then performs the new derived E-steps and M-steps iteratively until $\theta$ converges. Convergence analysis for EM was studied in literature and is out of the scope for this paper (Wu 1983).[1] Since each observed variable is binary, we can classify variables

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

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 $\theta^{(n)}$ (i.e., $a_i^c$, $b_i^c$ and $d_i^c$) based on Eq. (4). Algorithm 1 shows the pseudocode of OtO EM.

## 4 Accounting for variable 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.

### 4.1 Deriving the likelihood

In order to derive a likelihood function that considers constraints in the form of constraints 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 $g_1, \ldots, g_k$. Let $p(z_{g_1}, \ldots, z_{g_k})$ represent the joint probability distribution of the $k$ variables and let $\mathcal{Y}_g$ represent all possible combinations of values of $g_1, \ldots, g_k$. For example, when there are only two variables, $\mathcal{Y}_g = [(1, 1), (1, 0), (0, 1), (0, 0)]$. Note that, we assume that $p(z_{g_1}, \ldots, z_{g_k})$ is known or can be estimated from prior knowledge. The new likelihood function $L(\theta; X, Z)$ that considers the aforementioned constraints is:

$$L(\theta; X, Z) = \prod_{g \in G} p(X_g, Z_g|\theta) = \prod_{g \in G} p(Z_g) \times p(X_g|Z_g, \theta)$$
$$= \prod_{g \in G} \left\{ \sum_{g_1, \ldots, g_k \in \mathcal{Y}_g} p(z_{g_1}, \ldots, z_{g_k}) \prod_{i \in M} \prod_{j \in c_g} \alpha_{i,j} \right\} \tag{11}$$

where $\alpha_{i,j}$ is the same as defined in Eq. (10) and $c_g$ represents the set of variables belonging to the independent group $g$. Compared to our previous effort (Wang et al. 2012b), 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 Sect. 1.

### 4.2 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 observation 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 binary variable classification. The OtO+DV EM pseudocode is shown in Algorithm 2.

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

Initialize $\theta$ with random values between 0 and 1
**while** $\theta^{(n)}$ does not converge **do**
  **for** $j = 1 : N$ **do**
    compute $Z(n, j)$ as the *marginal distribution* of the joint probability as shown in Eq. (28)
  **end for**
  $\theta^{(n+1)} = \theta^{(n)}$
  **for** $i = 1 : M$ **do**
    compute $a_i^{(n+1)}, b_i^{(n+1)}, d_j^{(n+1)}$ based on Eq. (29)
    update $a_i^{(n)}, b_i^{(n)}, d_j^{(n)}$ with $a_i^{(n+1)}, b_i^{(n+1)}, d_j^{(n+1)}$ in $\theta^{(n+1)}$
  **end for**
  $t = t + 1$
**end while**
Let $Z_j^c =$ converged value of $Z(n, j)$
Let $a_i^c =$converged value of $a_i^{(n)}$; $b_i^c =$converged value of$b_i^{(n)}$; $d_i^c =$converged value of $d_j^{(n)}$ $j \in C_i$
**for** $j = 1 : N$ **do**
  **if** $Z_j^c \geq threshold$ **then**
    the value of $C_j$ 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

## 5 The analytic bound

In the previous section, we derived the OtO, DV, OtO+DV EM schemes to address the constraints on both sources and the observed variables. However, one important question remains: how to quantify the accuracy of the estimation results? In particular, we are interested in obtaining the confidence intervals; namely, the error bounds on the estimation parameters of our model for a given confidence level. In this section, we derive such bounds by using the Cramer–Rao lower bounds (CRLB) from estimation theory. We should note that the CRLBs derived here are assuming that enough sources are available so that the truth of the variable (or not) is known with full accuracy. As a result, the CRLBs are asymptotic results.

### 5.1 Deriving error bounds

We start with the derivation of Cramer–Rao lower bounds for our problem. The CRLB states the lower bounds of estimation variance that can be achieved by the MLE. By definition of CRLB, it is given by

$$CRLB = J^{-1} \tag{12}$$

where

$$J = E\left[\bigtriangledown_\theta \ln p(X|\theta)\ \bigtriangledown_\theta^H \ln p(X|\theta)\right] \tag{13}$$

where $J$ is the Fisher information of the estimation parameter, $\bigtriangledown_\theta = (\frac{\partial}{\partial a_1}, \ldots \frac{\partial}{\partial a_M}, \frac{\partial}{\partial b_1}, \ldots, \frac{\partial}{\partial b_M})^H$ and $H$ denotes the conjugate transpose operation.

In this subsection, we derive the asymptotic CRLBs for OtO EM, DV EM, and OtO+DV EM based on the assumption that the values of variables are correctly estimated by the EM algorithms. This is a reasonable assumption when the number of sources is enough (Wang et al. 2012a). We denote the log-likelihood function obtained under this assumption as $l_{em}(x;\theta)$.

We compute the Fisher Information Matrix from its definition. Note that CRLB should use the actual ground truth values of $a_i$ and $b_i$. However, due to the lack of ground truth in many real world applications, the MLE values of $a_i$ and $b_i$ are incorporated here as a means to approximate the expected variance. We can derive the representative element of Fisher Information Matrix from N variables as:

$$(J(\hat{\theta}_{MLE}))_{i,j} = \begin{cases} 0 & i \neq j \\ -E_X\left[\frac{\partial^2 l_{em}(x;a_i)}{\partial a_i^2}|_{a_i=\hat{a}_i^{MLE}}\right] & i = j \in [1, M] \\ -E_X\left[\frac{\partial^2 l_{em}(x;b_i)}{\partial b_i^2}|_{b_i=\hat{b}_i^{MLE}}\right] & i = j \in (M, 2M] \end{cases} \tag{14}$$

For the OtO EM, the log-likelihood function $l_{em}(x;\theta)$ can be written as follows:

$$\begin{aligned} l_{em}(x;\theta) = \sum_{j=1}^{N} \Bigg\{ z_j \times \Bigg[ \sum_{i\in\mathcal{S}_j} (S_i C_j \log a_i + (1 - S_i C_j)\log(1 - a_i) + \log d_j) \Bigg] \\ + (1 - z_j) \times \Bigg[ \sum_{i\in\mathcal{S}_j} \big(S_i C_j \log b_i + (1 - X_{ij})\log(1 - b_i) \\ + \log(1 - d_j)\big) \Bigg] \Bigg\} \end{aligned} \tag{15}$$

where $z_j$ is the converged value of $Z(n, j)$ in (23).

Substituting the log-likelihood function in Eq. (15) into Eq. (14), the CRLB of OtO EM (i.e., the inverse of the Fisher Information Matrix) can be written as:

$$CRLB^{OtO} = \begin{cases} 0 & i \neq j \\ \frac{\hat{a}_i^{MLE} \times (1-\hat{a}_i^{MLE})}{\mathcal{C}_i \times d_i} & i = j \in [1, M] \\ \frac{\hat{b}_i^{MLE} \times (1-\hat{b}_i^{MLE})}{\mathcal{C}_i \times (1-d_i)} & i = j \in (M, 2M] \end{cases} \tag{16}$$

where $\mathcal{C}_i$ is the set of variables that $S_i$ observed and $d_i$ is defined in Eq. (4). The $\hat{a}_i^{MLE}, \hat{b}_i^{MLE}$ are derived in Derivation of the E-step and M-step of OtO EM in Appendix and the results are shown by Eq. (24).

Following similar derivation steps, we can also derive the CRLB of DV EM as:

$$CRLB^{DV} = \begin{cases} 0 & i \neq j \\ \frac{\hat{a}_i^{MLE} \times (1-\hat{a}_i^{MLE})}{N \times d} & i = j \in [1, M] \\ \frac{\hat{b}_i^{MLE} \times (1-\hat{b}_i^{MLE})}{N \times (1-d)} & i = j \in (M, 2M] \end{cases} \tag{17}$$

where the $\hat{a}_i^{MLE}, \hat{b}_i^{MLE}$ are derived in Derivation of E-step and M-step of DV and OtO+DV EM in Appendix and the results are shown by Eq. (27).

Finally, the CRLB of OtO+DV EM can derived as:

$$CRLB^{OtO+DV} = \begin{cases} 0 & i \neq j \\ \frac{\hat{a}_i^{MLE} \times (1-\hat{a}_i^{MLE})}{\mathcal{C}_i \times d_i} & i = j \in [1, M] \\ \frac{\hat{b}_i^{MLE} \times (1-\hat{b}_i^{MLE})}{\mathcal{C}_i \times (1-d_i)} & i = j \in (M, 2M] \end{cases} \tag{18}$$

where the $\hat{a}_i^{MLE}, \hat{b}_i^{MLE}$ are derived in The confidence interval in Appendix and the results are shown by Eq. (29).

### 5.2 The confidence interval

In this subsection, we show that the confidence interval of source reliability (i.e., the probability a source $S_i$ makes a correct observation) can be obtained by using the CRLB we just derived and the asymptotic normality of the MLE.

One of the attractive asymptotic properties about maximum likelihood estimator is called *asymptotic normality*: The MLE estimator is asymptotically distributed with Gaussian behavior as the data sample size goes up, in particular (Casella and Berger 2002):

$$\left(\hat{\theta}_{MLE} - \theta_0\right) \xrightarrow{d} N\left(0, J^{-1}\left(\hat{\theta}_{MLE}\right)\right) \tag{19}$$

where $J$ is the Fisher Information Matrix computed from all samples, $\theta_0$ and $\hat{\theta}_{MLE}$ are the true value and the MLE of the parameter $\theta$ respectively. The Fisher information at the MLE is used to estimate its true (but unknown) value (Hogg and Craig 1995).

Following the asymptotic normality of the maximum likelihood estimator (Cramer 1946), the error of the corresponding estimation on $\theta$ follows a normal distribution

with zero mean and the covariance matrix given by the CRLB we derived in the previous subsection. The variance of estimation error on parameter $a_i$ is denoted as $var(\hat{a}_i^{MLE})$. For a problem with sufficient M and N (i.e., under asymptotic condition), $(\hat{t}_i^{MLE} - t_i^0)$ also follows a norm distribution with 0 mean and variance given by:

$$var\left(\hat{t}_i^{MLE}\right) = \left(\frac{d_i}{s_i}\right)^2 var\left(\hat{a}_i^{MLE}\right) \tag{20}$$

Thus, the confidence interval that can be used to quantify the source reliability (i.e., $t_i$) is given by the following:

$$\left(\hat{t}_i^{MLE} - c_p\sqrt{var(\hat{t}_i^{MLE})}, \hat{t}_i^{MLE} + c_p\sqrt{var(\hat{t}_i^{MLE})}\right) \tag{21}$$

where $c_p$ is the standard score (z-score) of the confidence level $p$. For example, for the 95 % confidence level, $c_p = 1.96$.

## 6 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 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 (Wang et al. 2012b) (regular EM) through both a real world social sensing application and extensive simulation studies. We also evaluated the performance of the analytic bounds derived in the previous section.

### 6.1 Real world evaluation

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 (Hu et al. 2013), 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 s. 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 sensing observations 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.

Specifically, in our experiment, if a vehicle waits at a location for 15–90 s, the application concludes that it is stopped at a traffic light and issues a traffic-light observation (i.e., an observation that a traffic light is present at that location and bearing). Similarly if it waits for 2–10 s, it concludes that it is at a stop sign and issues a stop-sign observation (i.e., an observation that a stop sign is present at that location and

bearing). If the vehicle stops for less than 2 s, for 10–15 s, or for more than 90 s, no observation is made. Observations were reported by each participant to a central data collection point.

Clearly the observations 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 sources. One should note that location granularity here is of the order of half a city block. This ensures that stop sign and traffic light observations are attributed to the correct intersections. Most GPS devices easily attain such granularity. Therefore, we 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 h of driving) were collected. A total of 4865 observations 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 $S_i C_j$ of the observation matrix were set according to the reported observations extracted from each source vehicle.

We observed that traffic lights at an intersection are always present in all directions. Hence, when processing traffic light observations, 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 observations regarding stop signs the bearing is important. We bin bearing into four main directions. A different Boolean variable is created for each direction.

### *6.1.1 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.

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

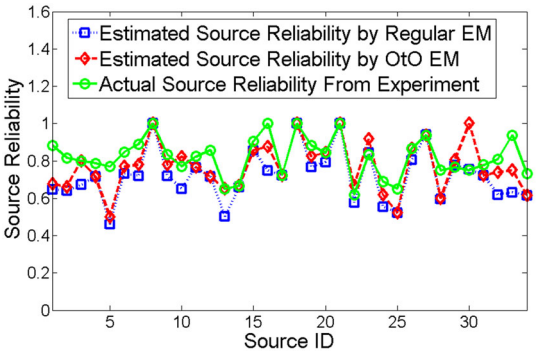



**Fig. 1** Source reliability estimation of OtO EM in the case of traffic lights

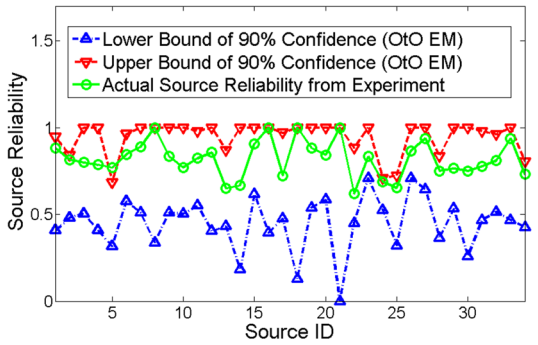



**Fig. 2** Source reliability bounds of OtO EM in the case of traffic lights

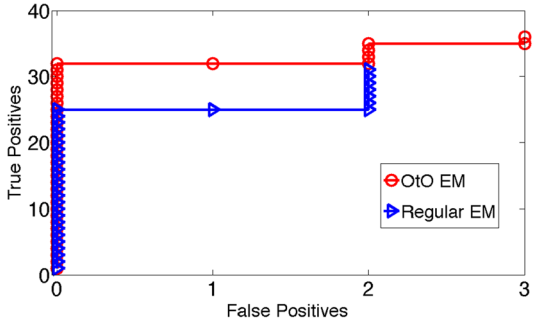



**Fig. 3** True and false positives of OtO EM versus regular EM in the case of traffic lights

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).

Figure 2 shows the 90 % confidence bounds on the source reliability estimation by the OtO EM as we derived in Sect. 5. We observed the OtO EM scheme has only one outlier out of 34 sources, which matches well with the definition of the confidence bounds defined at this confidence level.

Next, we explore the accuracy of identifying traffic lights by the new scheme. We plotted the true positives and false positives of the OtO EM scheme and the regular EM scheme for the locations they identified as traffic lights. The results are shown in Fig. 3. We observed the OtO EM scheme outperforms the regular EM by finding more

**Table 1** Performance comparison between regular EM versus OtO EM in case of traffic lights

| | 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 |

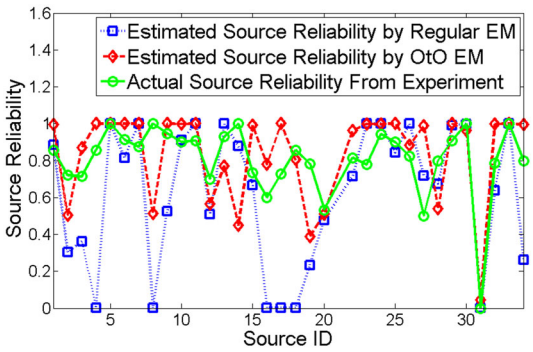



**Fig. 4** Source reliability estimation of OtO EM in the case of stop signs

true positives at the same false positives. In particular, 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 1.

We repeated the above experiments for stop sign identification and observed that the OtO EM scheme achieves a more significant performance gain in both source 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 observations 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 observations 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 shows the 90 % confidence bounds on the source reliability estimation by the OtO EM in the case of stop signs. We observed that the OtO EM scheme has only

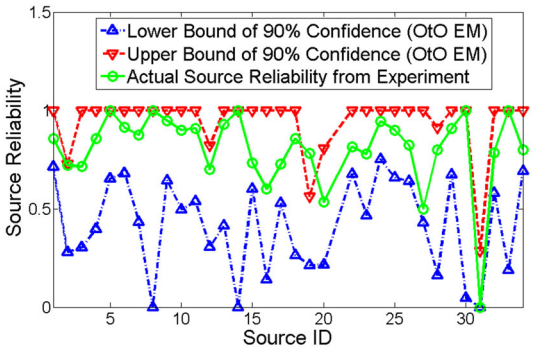



**Fig. 5** Source reliability bounds of OtO EM in the case of stop signs

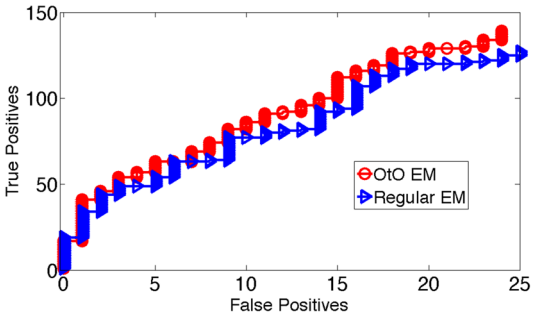



**Fig. 6** True and false positives of OtO EM versus regular EM in the case of stop signs

one outlier out of 34 sources. This again verifies the correctness of the confidence bounds we derived earlier.

Figure 6 show the true positives and false postives in recognizing stop signs. We observe the OtO EM scheme outperforms the regular EM scheme. In particular, the OtO EM 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 Table 2. To further investigate the effects of data sparsity on different schemes, we repeat the above experiments using only 75 % of the observations we collected. Results are also reported in Table 2.

### *6.1.2 Dependent variables*

In this subsection, we evaluated our extensions that consider constraints on observed variables (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

**Table 2** Performance comparison of regular EM, OtO EM, DV EM and DV+OtO EM in case of stop signs

| | 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 3** Distribution of stop signs in opposite directions

| 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 |

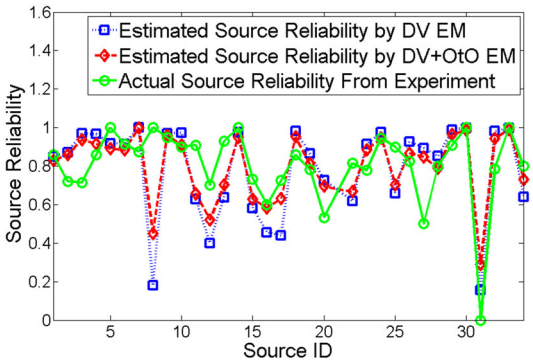



**Fig. 7** Source reliability estimation of DV and DV+OtO EM in the case of stop signs

co-existence of (two) stop signs in opposite directions at an intersection was computed. It is presented in Table 3, 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).

Figures 8 and 9 show the 90 % confidence bounds on the source reliability estimation by the DV EM and DV+OtO EM respectively. We observed the DV EM scheme has two outliers out of 34 sources while DV+OtO EM scheme has no outlier. These results

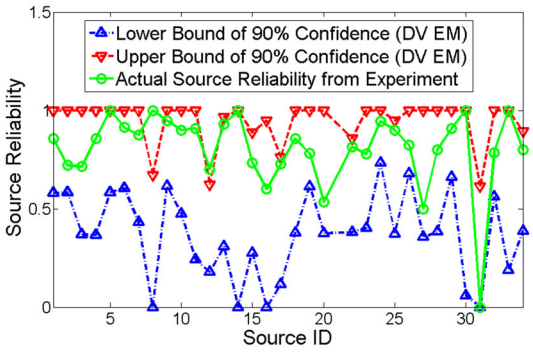



**Fig. 8** Source reliability bounds of DV EM in the case of stop signs

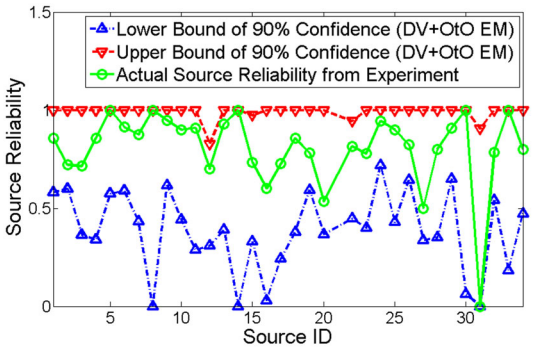



**Fig. 9** Source reliability bounds of DV+OtO EM in the case of stop signs

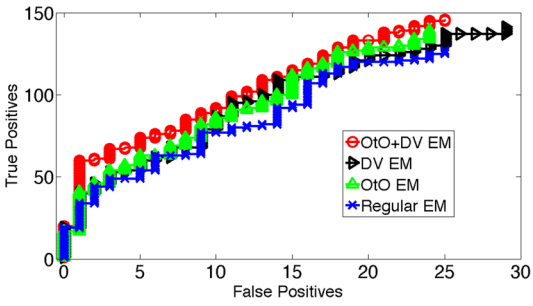



**Fig. 10** ROC curves of OtO, DV, OtO+DV EM versus regular EM in the case of stop signs

are encouraging. They verified the correctness of the confidence bounds we derived to quantify the accuracy of the source reliability estimation by the new EM schemes developed in this paper.

The true positives and false positives of DV and DV+OtO EM for stop signs are shown in Fig. 10. Observe that the DV EM scheme finds 14 more correct stop sign locations than the regular EM scheme. 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 results are given in Table 2.

### 6.2 Simulation study

In this section, we continued our evaluation of the new schemes developed in this paper through extensive simulation studies to explore different problem dimensions. To that end, we built a simulator in Matlab 7.14.0 that generates a random number of sources and binary variables.[2] A random probability $t_i$ is assigned to each source $S_i$ representing his/her reliability (i.e., the ground truth probability that they report correct observations). For each source $S_i$, $L_i$ observations are generated. Each observation has a probability $t_i$ of being true (i.e., reporting a the value of a variable as true correctly) and a probability $1 - t_i$ of being false (reporting the value of a variable as true when it is not). We let $t_i$ be uniformly distributed between 0.5 and 1 in our experiments.[3] For initialization, the initial values of source reliability (i.e., $t_i$) in the evaluated schemes are set to the mean value of its definition range.

We compared the new schemes presented in this paper (i.e., OtO EM, DV EM, OtO+DV EM) with the regular EM scheme, which was reported to beat four other state-of-the-art baselines (Wang et al. 2012b). To evaluate the performance of different schemes, we studied three metrics: (i) estimation error of source reliability; (ii) the fraction of misclassified variables; (iii) the correctness of the derived confidence bounds.

#### *6.2.1 OtO EM performance study*

In the first set of experiments, we studied the performance of the OtO EM scheme. In the experiment, the number of reported variables was fixed at 2000, of which 1000 variable were of true values and 1000 were of false values. The average number of observations per source was set to 100. The number of sources was varied from 30 to 120. For this set of experiments, we assumed variables are all independent. Reported results are averaged over 100 random source reliability distributions. We compare the OtO EM with regular EM under four scenarios where the fraction of observable variables is different. The fraction of observable variables is defined as the fraction of variables that a source has opportunity to observe. We also add an additional baseline as OtO EM + 50 % Uncertainty. This baseline is the same as the OtO EM scheme only except the sources now have 50 % probability to mis-identify the variables they do not have an opportunity to observe as observable. Results are shown in Figs. 11, 12, and 13. Figure 11 shows the results of the source reliability estimation error. We observed that the OtO EM consistently performed the best in all four scenarios. Also note that the performance gain achieved by OtO EM is larger when the fraction of observable variables is lower, which is intuitive. Figure 12 shows the results of variable classification accuracy. We observed that the OtO EM classifies more variables correctly compared to the regular EM and OtO EM+50 % Uncertainty by having a more accurate knowledge of "opportunity to observe" of sources. Figure 13

[2] As stated in our application model, sources never report a variable to be false (e.g., cars never reported the absence of traffic lights).

[3] In principle, there is no incentive for a source to lie more than 50 % of the time, since negating their statements would then give a more accurate truth.

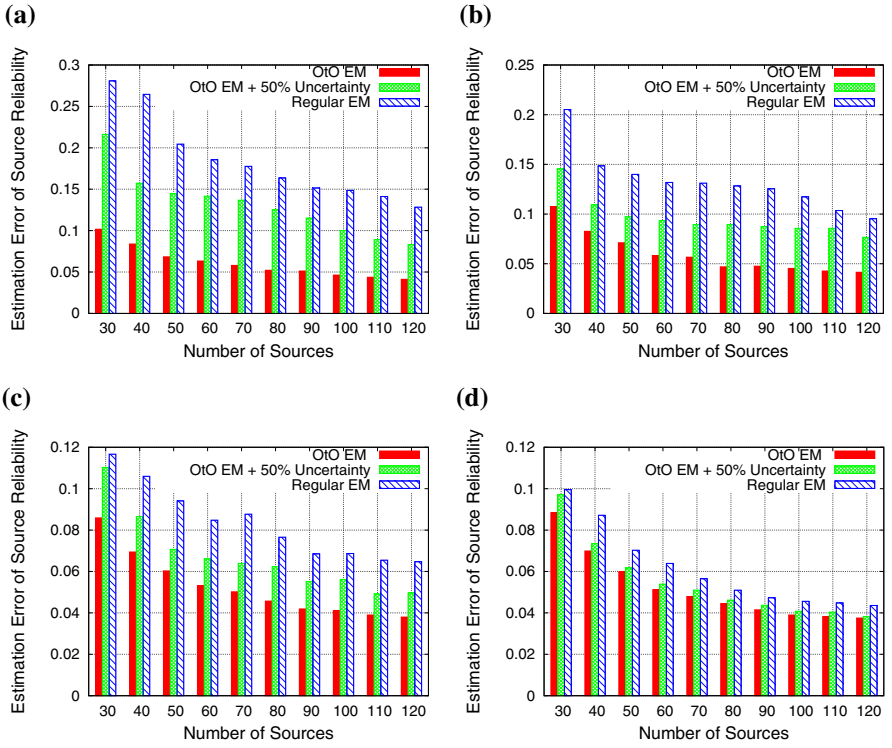



**Fig. 11** Source reliability estimation error of OtO EM versus regular EM. **a** Fraction of observable variables = 0.2. **b** Fraction of observable variables = 0.4. **c** Fraction of observable variables = 0.6. **d** Fraction of observable variables = 0.8

shows the fraction of sources whose reliability is correctly bounded by the 90 % confidence bounds computed in Sect. 5. We observed that the source reliability of the OtO EM scheme is correctly bounded by the corresponding confidence bounds in all scenarios while the bounds for the regular EM failed to be accurate when the fraction of observable variables is low.

Additionally, we also studied the performance trade-off between estimation variance and bias for the OtO EM scheme. We found the estimation bias is more significant when the number of sources in the system is small and showed in our previous work that the actual CRLBs track the estimation variance better than the asymptotic bounds under such conditions (Wang et al. 2013c). Hence we computed the actual CRLB and estimation variance for both OtO EM and regular EM for a small number of sources. The experiment setup is the same as before. We now varied the number of sources from 5 to 20. The fraction of observable variables is set to 0.5. The results are averaged over 50 experiments and shown in Fig. 14. We observe that the OtO EM has a larger CRLB and much smaller estimation bias compared to the regular EM scheme. The results demonstrate that the lower CRLB for regular EM does not translate to better estimation performance due to the bias. We also observe that there is some bias for OtO EM scheme when the number of sources is very small. This is because: (i) the MLE

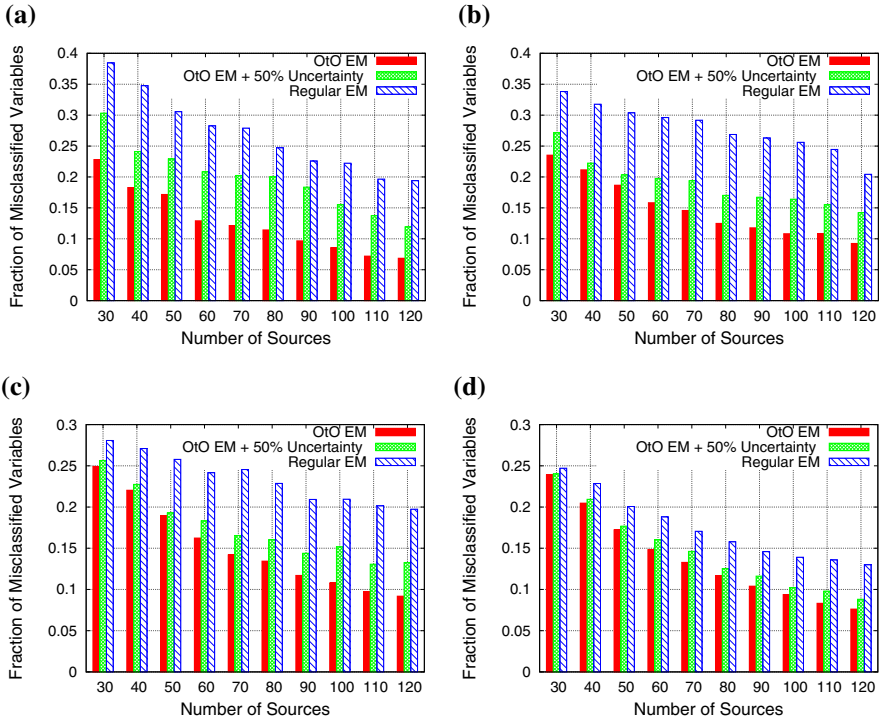



**Fig. 12** Fraction of misclassified variables of OtO EM versus regular EM. **a** Fraction of observable variables = 0.2. **b** Fraction of observable variables = 0.4. **c** Fraction of observable variables = 0.6. **d** Fraction of observable variables = 0.8

is biased on those points due to the small dataset; (ii) the number of variables made per source is not large enough to completely reflect the source reliability accuracy resolution.

### *6.2.2 DV EM performance study*

In the second set of experiments, we studied the performance of the DV EM scheme. The experiment setup is similar as the first one. The differences are (i) we assume sources have opportunity to observe all variables; (ii) variables are divided into independent groups and variables within each independent groups are dependent. For simplicity, we assumed all groups are of the same size and the variables within each group are fully correlated (i.e., the probabilities of correlated variables to have the same value are equal and add up to 1). Reported results are averaged over 100 random source reliability distributions. We compare the DV EM with regular EM under four scenarios where the fraction of dependent variables is different. The fraction of dependent variables is the fraction of variables that belong to independent group with more than one variable. In each scenario, we also vary the number of variables in each independent group (i.e., independent group size). Results are shown in Figs. 15,

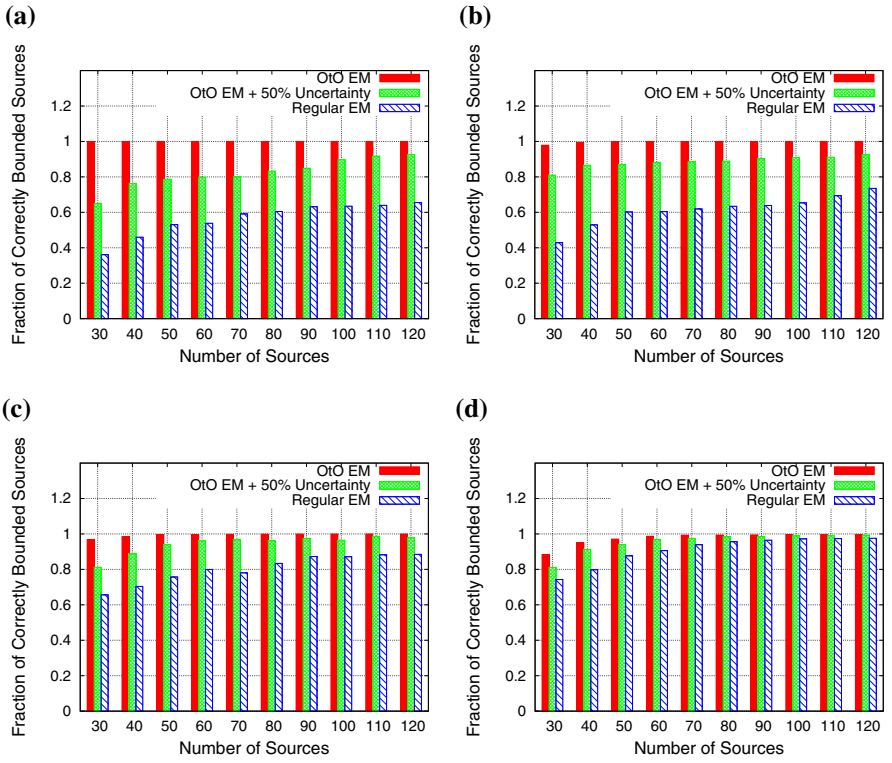



**Fig. 13** Fraction of correctly bounded sources of OtO EM versus regular EM. **a** Fraction of observable variables = 0.2. **b** Fraction of observable variables = 0.4. **c** Fraction of observable variables = 0.6. **d** Fraction of observable variables = 0.8

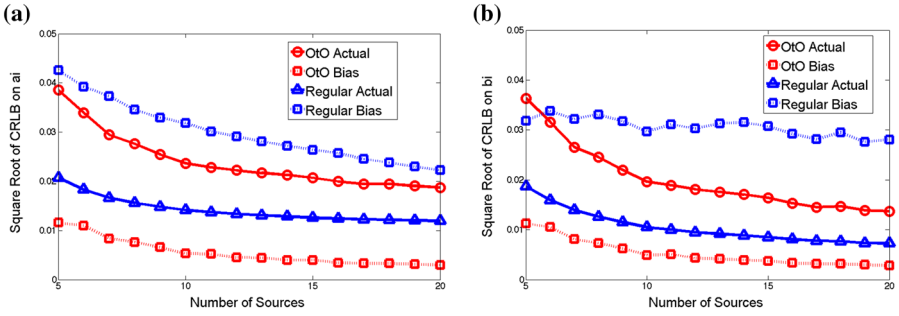



**Fig. 14** Tradeoff between estimation variance and bias of OtO versus regular EM. **a** CRLB and bias on $a_i$. **b** CRLB and bias on $b_i$

16, and 17. Figure 15 shows the results of the source reliability estimation error. We observed that the DV EM performed better than the regular EM in all four scenarios. Also note that the performance gain of DV EM is larger when the independent group size is larger. This is because more correlations between variables can help the DV EM scheme to better infer correctness of all dependent variables. Figure 16 shows

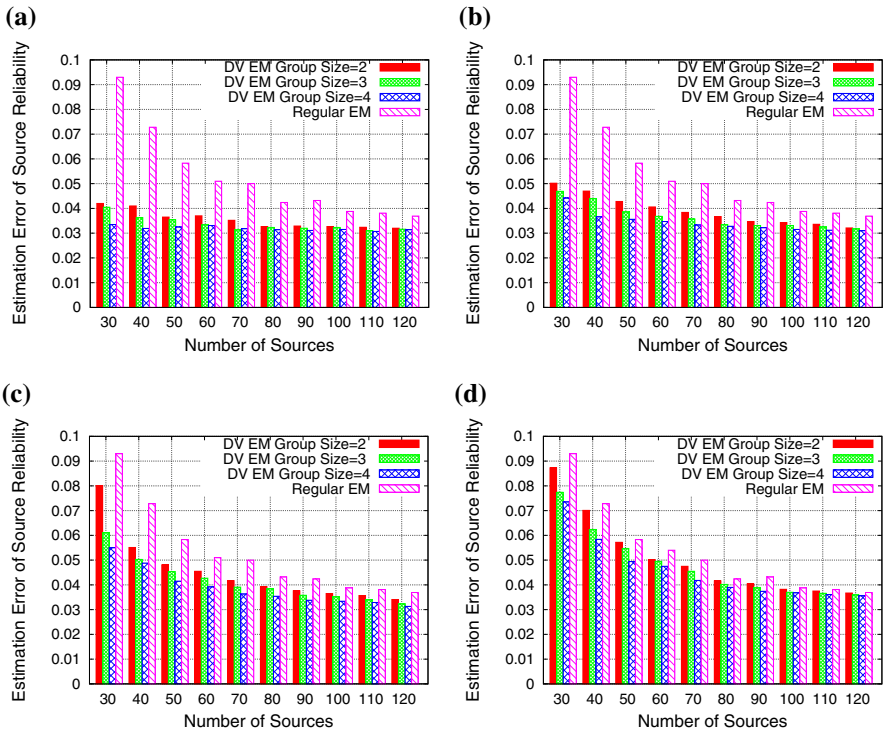



**Fig. 15** Source reliability estimation error of DV EM versus regular EM. **a** Fraction of dependent variables = 1. **b** Fraction of dependent variables = 0.8. **c** Fraction of dependent variables = 0.5. **d** Fraction of dependent variables = 0.2

the results of variable classification accuracy. We observed that the DV EM classifies more variables correctly compared to the regular EM by appropriately handling the constraint between variables. Figure 17 shows the fraction of sources whose reliability is correctly bounded by the 90 % confidence bounds. We observed that the source reliability of the DV EM scheme is correctly bounded by the corresponding confidence bounds in all scenarios while the bounds for the regular EM failed to be accurate when the number of sources in the system is small.

Additionally, we also studied the convergence performance of the DV EM scheme. We derived the actual CRLBs for the regular EM scheme in (Wang et al. 2013c). We found it is non-trivial to derive the actual CRLBs for the DV EM under arbitrary variable constraints and decided to leave such derivations for a follow-up work. However, it is possible to compute the actual CRLB of DV EM for the special case where the variables in each independent group are fully correlated. Therefore, we compared the actual CRLBs of DV EM and regular EM scheme (under the condition of fully correlated variables in each group) with the asymptotic bounds we derived earlier. The experiment setup is the same as before. The fraction of dependent variables is set to 1 and the size of independent group is set to 2. We varied the number of sources from 5 to 20. The results are averaged over 50 experiments and shown in Fig. 18. We observe

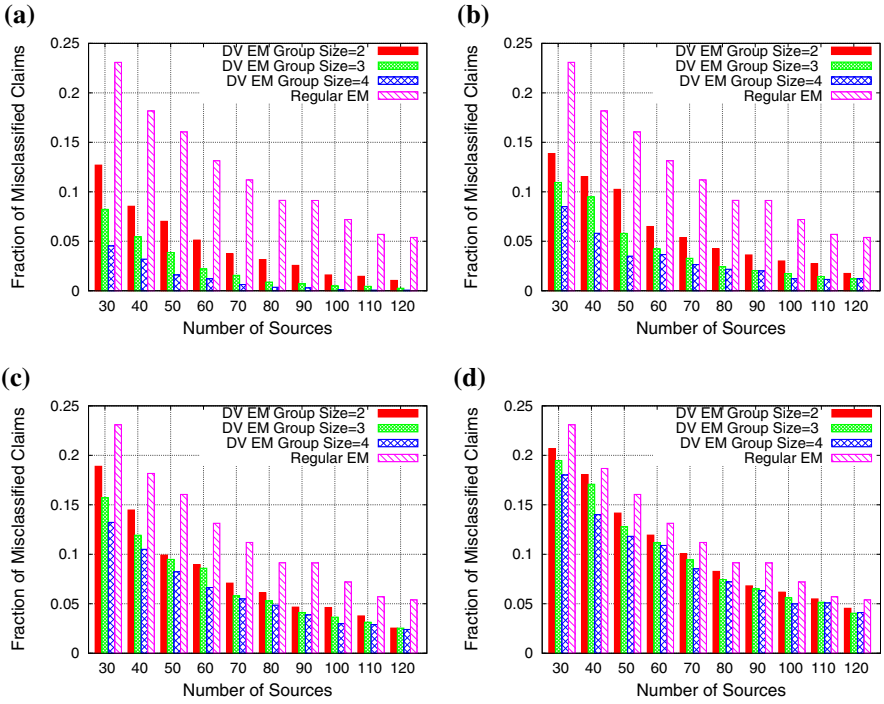



**Fig. 16** Fraction of misclassified variables of DV EM versus regular EM. **a** Fraction of dependent variables = 1. **b** Fraction of dependent variables = 0.8. **c** Fraction of dependent variables = 0.5. **d** Fraction of dependent variables = 0.2

that the DV EM has smaller CRLBs (for both $a_i$ and $b_i$) and converges faster to the asymptotic bounds compared to the regular EM scheme.

### *6.2.3 OtO+DV EM performance study*

In the third set of experiments, we studied the performance of the OtO+DV EM scheme in comparison with OtO EM, DV EM and regular EM scheme. The experiment setup is the same as before. However, we assumed in this set of experiments: (i) sources have opportunity to observe only a fraction of all variables; (ii) a fraction of variables are dependent and the remaining ones are independent. Reported results are averaged over 100 random source reliability distributions. Results are shown in Figs. 19, 20, and 21. Figure 19 shows the results of different schemes by varying the number of sources in the system. In this experiment, we set both the fraction of observable variables and the fraction of dependent variables to be 0.8. The number of sources was varied from 30 to 120. We observed that the OtO + DV EM performed the best compared to other baselines in all evaluation metrics. Also note that the performance of all schemes improves as the number of sources increases. Figure 20 shows the results of different schemes by varying the fraction of observable variables. The number of sources was set to 30 and the fraction of dependent variables was set to 0.8. We varied

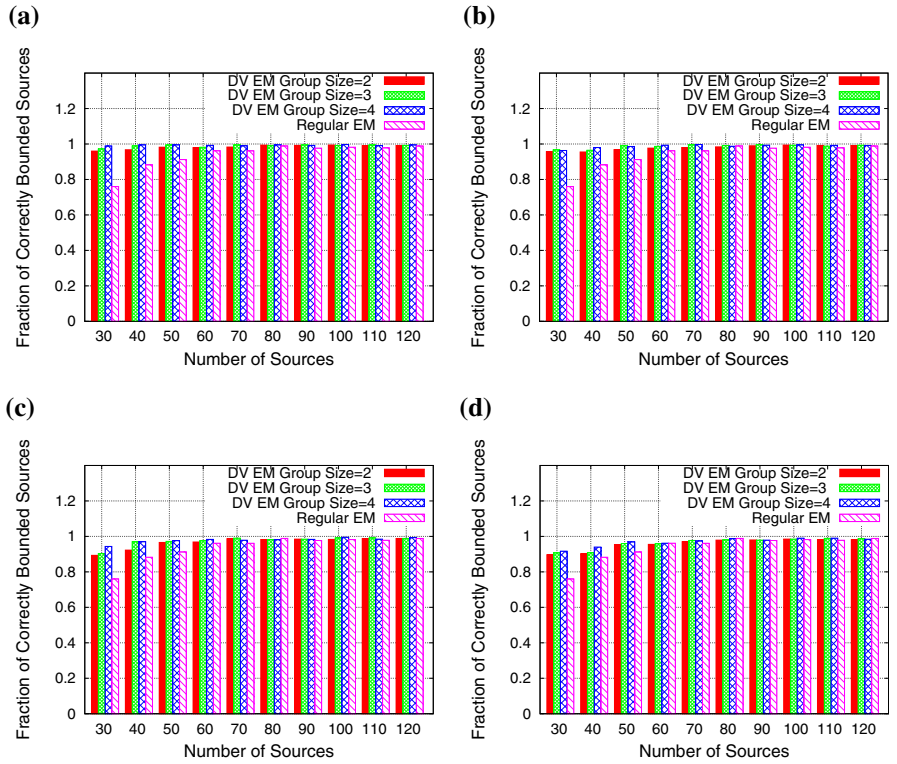



**Fig. 17** Fraction of correctly bounded sources of DV EM versus regular EM. **a** Fraction of dependent variables = 1. **b** Fraction of dependent variables = 0.8. **c** Fraction of dependent variables = 0.5. **d** Fraction of dependent variables = 0.2

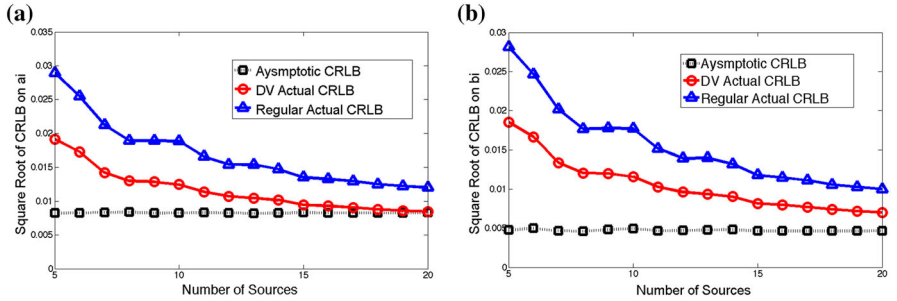



**Fig. 18** Convergence of actual CRLB of DV versus regular EM. **a** CRLB on $a_i$. **b** CRLB on $b_i$

the fraction of observable variables from 0.1 to 1. We observed that the OtO+DV EM continues to have the best performance among all schemes under comparison. We also noted the performance of the schemes that ignore "opportunity to observe" (i.e., DV EM and regular EM) becomes worse as the fraction of observable variables in the system decreases. Figure 21 shows the results of different schemes by varying the fraction of dependent variables. The number of sources was kept the same as the

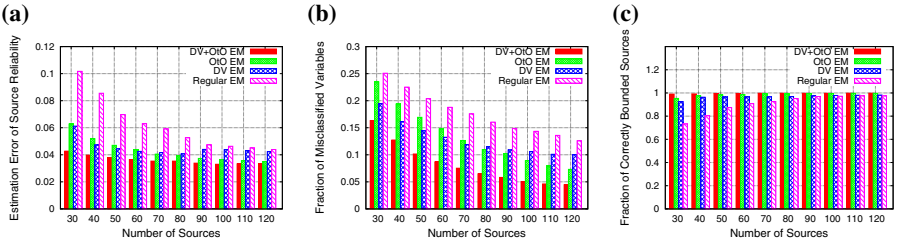



**Fig. 19** OTO+DV EM, OTO EM, DV EM, and regular EM versus varying the number of sources. **a** Estimation error of source reliability. **b** Fraction of misclassified variables. **c** Fraction of correctly bounded sources

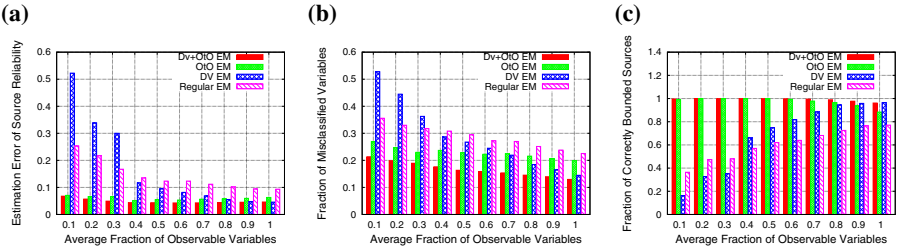



**Fig. 20** OTO+DV EM, OTO EM, DV EM, and regular EM versus varying the fraction of observable variables. **a** Estimation error of source reliability. **b** Fraction of misclassified variables. **c** Fraction of correctly bounded sources

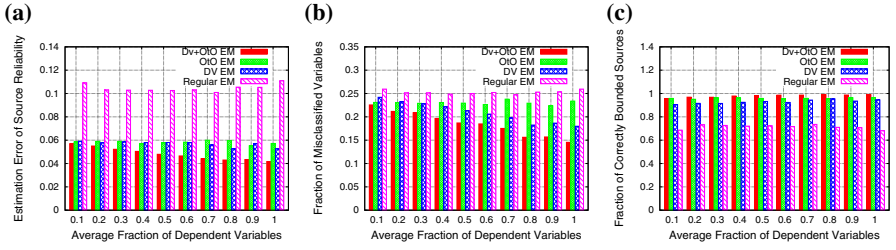



**Fig. 21** OTO+DV EM, OTO EM, DV EM, and regular EM versus varying the fraction of dependent variables. **a** Estimation error of source reliability. **b** Fraction of misclassified variables. **c** Fraction of correctly bounded sources

previous experiment and the fraction of observable variables was set to 0.8. We varied the fraction of dependent variables from 0.1 to 1. We observed that the OtO+DV EM is still the best performed scheme compared to other baselines. Also note that the performance of the schemes that take variable constraint into account (i.e., OtO+DV EM and DV EM) improves as the fraction of dependent variables in the system increases.

## 7 Discussion and limitations

Motivated by the need to address data reliability challenges in emerging cyber-physical systems (with humans-in-the-loop), this paper presented a MLE framework for exploiting the physical world constraints (i.e., source locations and observed vari-

able constraints) to improve the reliability of social sensing. Some limitations exist that offer directions for future work.

First, we do not explicitly address *uncertainties in reported locations* and the observed variables in our model are assumed to be *time invariant*. This is mainly because our current application involves the detection of fixed infrastructure (e.g., locations of stop signs and traffic lights where the localization accuracy of the GPS is sufficient). Time is also less relevant in such context. Hence, the source constraint is only a function of source location, and observed variable constraints are not likely to change over time. In systems where the state of the environment may change over time, when we consider the source constraints, 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 constraints, it is crucial that constraints of observed variables remain stable within a given time interval and we have an efficient way to quickly update our estimation on such constraints as time goes by. More recently, we have developed an extended MLE framework to explicitly handle *time variant variables* in our model (Wang et al. 2014b). The intuition of the new approach is that we could model the correlations between different states of a time variant variable in a similar way as we model observed variable constraints, which is discussed in this paper. In these applications, the localization error is not an issue, but in general such errors can be problematic, e.g., when collecting information from social media such as Twitter. Future work needs to consider such errors in the more general setting.

Second, we assume sources will only report observations for the places they have been to (e.g., cars only generate stop sign observations on the streets their GPS traces covered). Hence, it makes sense to "penalize" sources for not making observations for some clearly observable variables based on their locations. However, many other factors might also influence the opportunity of users to generate observations 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 (He et al. 2011) and social gaming systems (de Valmaseda et al. 2013). 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.

Finally, 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.

## 8 Related work

This work broadly falls into the area of addressing correctness challenges in cyber-physical systems. Significant prior advances were made in addressing timing correctness and functional correctness of cyber-physical systems (Liu and Layland 1973; Park et al. 1996; Pandya and Malek 1998; Mok and Chen 1997; Strosnider et al. 1995; Sprunt et al. 1989; Lin and Tarng 1991; Cook et al. 2005, 2006; Alur et al. 1995; Saeedloei and Gupta 2011). In real time community, a large number of literature centered around developing various scheduling policies and deriving corresponding utilization bounds to address timing correctness. A good survey of real-time scheduling policies can be found in (Sha et al. 2004). Liu and Layland presented the first utilization bound for periodic tasks on a single processor (Liu and Layland 1973). This work is followed by a plethora of work to improve the Liu and Layland bound in different dimensions such as run-time extension (Park et al. 1996), fault-tolerance extension (Pandya and Malek 1998), multi-frame periodic model extension (Mok and Chen 1997). Several algorithms have also been developed to derive the utilization bounds for aperiodic tasks (Strosnider et al. 1995; Sprunt et al. 1989; Lin and Tarng 1991). The functional correctness in CPS mainly refers to correctness of program logic and system modeling (Sha et al. 2009; Rajkumar et al. 2010). Useful results and tools have been recently developed for software verification and program analysis in cyber-physical and hybrid systems (Cook et al. 2005, 2006). Formalism based methods have also been developed to study the modeling correctness of CPS (Alur et al. 1995; Saeedloei and Gupta 2011). In contrast, this paper investigates *data correctness* challenges, which is motivated by cyber-physical applications with humans-in-the-loop; specifically the rise of applications that exploit social sensing.

Human-in-the-loop cyber-physical systems (HiLCPSs) incorporate a challenging and promising class of CPS applications that augment and facilitate human interaction with the physical world (Schirner et al. 2013). Some examples of these applications include energy management (Lu et al. 2010), health care (Kay et al. 2012), automobile systems (Ganti et al. 2010), and disaster response (Uddin et al. 2011). Many interesting research challenges have been studied in HiLCPSs applications (Munir et al. 2013). For example, Wolpaw et al. developed a non-invasive brain computer interface (BCI) to efficiently measure electric potential on the scalp for the infererence of human's intent (Wolpaw and Wolpaw 2012). Lu et al. designed a smart thermostat system by leavening hidden markov model to model occupancy and sleep pattern of the residents in a home for energy savings (Lu et al. 2010). The work in this paper is complementary to the work mentioned above. We focused on addressing the *data reliability* problems in HiLCPSs where humans play the role of sensors or sensor carriers

and where the reliability of data sources and the collected data is in general unknown a priori.

Social sensing is made possible by 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). In social sensing applications, humans play a key role in data collection by acting as sensor carriers (Lane et al. 2008) (e.g., opportunistic sensing), sensor operators (Burke et al. 2006) (e.g., participatory sensing) or sensor themselves. An early overview of social sensing applications is described in (Abdelzaher et al. 2007). Examples of early systems include CenWits (Huang et al. 2005), CarTel (Hull et al. 2006), BikeNet (Eisenman et al. 2007), and CabSense.[4] Recent work explored privacy (Pham et al. 2010), energy-efficient context sensing (Nath 2012), and social interaction aspects (Rachuri et al. 2011).

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 ascertaining correctness of data from unreliable sources (Kleinberg 1999; Yin et al. 2008; Pasternack and Roth 2010). More recent work on fact-finding came up with new algorithms by leveraging techniques in statistics and estimation theory (Zhao et al. 2012; Wang et al. 2012b, 2013a, c, 2014a). Zhao et al. (2012) 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. (2012b) proposed a MLE framework that offers a joint estimation on source reliability and variable classification based on a set of general simplifying assumptions. In their following work, Wang et al. further extended their framework to handle streaming data (Wang et al. 2013a) and source dependency (Wang et al. 2014a). The approach was compared to several state-of-the-art previous fact-finders and was shown to outperform them in estimation accuracy (Wang et al. 2012b). Accordingly, we only compare our new extensions to the winning approach from prior art. The accuracy of the MLE approach has been quantified in Wang et al. (2013c). However, the derivation of such accuracy bounds are based on the assumptions that sources have opportunities to observe the underlying events of *all variables* and variables are *independent*. In contrast, we derived new accuracy bounds in this paper that relaxed the above two assumptions. In other words, our bounds accommodated a source's opportunity to observe a subset of variables and constraint between different variables.

Finally, physical correlations 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 dependency to reduce data transmission needs. Compression and coding schemes were proposed to reduce the data redundancy in the space domain (Scaglione and Servetto 2002; Xu and Lee 2006). Temporal correlations were exploited to reduce network load while offering compression quality guarantees (Guitton et al. 2007; Ali et al. 2011). The novelty of our work lies in incorporating the constraints from the physical world into a framework for *improving estimation accuracy* as opposed to *reducing*

[4] CabSense. http://www.cabsense.com.

*resource cost*. The underlying insight is the same: knowledge of physical 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.

## 9 Conclusion

This paper presented a framework and new analytic bounds for incorporating source and observed variable constraints that arise from physical knowledge (of source locations and observed variable correlations) into maximum-likelihood analysis to improve the accuracy of social sensing in cyber-physical applications with humans-in-the-loop. The problem addressed was one of jointly assessing the values of observed variables 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 and analytic bounds was evaluated through both a real world social sensing application and extensive simulation studies. Results show a significant reduction in estimation error of both source reliability and variable classification as well as the correctness of derived analytic bounds thanks to the exploitation of physical constraints.

**Acknowledgments** Research reported in this paper was sponsored by National Science Foundation under Grant No. IIS-1447795 and Army Research Laboratory under Cooperative Agreement W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

## Appendix

### 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 Eq. (7) of Expectation Maximization. The E-step can be derived as follows:

$$\begin{aligned} Q\left(\theta|\theta^{(n)}\right) &= E_{Z|X,\theta^{(n)}}[\log L(\theta; X, Z)] \\ &= \sum_{j=1}^{N} \Bigg\{ p(z_j = 1|X_j, \theta^{(n)}) \times \sum_{i \in \mathcal{S}_j} (\log \alpha_{i,j} + \log d_j) \\ &\quad + p(z_j = 0|X_j, \theta^{(n)}) \times \sum_{i \in \mathcal{S}_j} (\log \alpha_{i,j} + \log(1 - d_j)) \Bigg\} \end{aligned} \tag{22}$$

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

$$\begin{aligned} Z(n, j) &= p(z_j = 1|X_j, \theta^{(n)}) \\ &= \frac{p(z_j = 1; X_j, \theta^{(n)})}{p(X_j, \theta^{(n)})} \\ &= \frac{p(X_j, \theta^{(n)}|z_j = 1)p(z_j = 1)}{p(X_j, \theta^{(n)}|z_j = 1)p(z_j = 1) + p(X_j, \theta^{(n)}|z_j = 0)p(z_j = 0)} \\ &= \frac{\prod_{i \in \mathcal{S}_j} \alpha_{i,j} \times d_j^{(n)}}{\prod_{i \in \mathcal{S}_j} \alpha_{i,j} \times d_j^{(n)} + \prod_{i \in \mathcal{S}_j} \alpha_{i,j} \times (1 - d_j^{(n)})} \end{aligned} \tag{23}$$

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 $\theta^*$ (i.e., $a_1^*, a_2^*, \ldots, a_M^*; b_1^*, b_2^*, \ldots, b_M^*; d_1^*, d_2^*, \ldots, d_N^*$) that maximizes the $Q\left(\theta|\theta^{(n)}\right)$ function in each iteration and is used as the $\theta^{(n+1)}$ of the next iteration.

$$\begin{aligned} a_i^{(n+1)} &= a_i^* = \frac{\sum_{j \in SJ_i} Z(n, j)}{\sum_{j \in \mathcal{C}_i} Z(n, j)} \\ b_i^{(n+1)} &= b_i^* = \frac{\sum_{j \in SJ_i} (1 - Z(n, j))}{\sum_{j \in \mathcal{C}_i} (1 - Z(n, j))} \\ d_j^{t+1} &= d_j^* = Z(n, j) \\ d_i^* &= \frac{\sum_{j \in \mathcal{C}_i} Z(n, j)}{|\mathcal{O}_i|} \end{aligned} \tag{24}$$

where $\mathcal{O}_i$ is set of variables source $S_i$ observes according to the knowledge matrix $SK$ and $Z(n, j)$ is defined in Eq. (23). $SJ_i$ is the set of variables the source $S_i$ actually reports in the observation 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.

### 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 Eq. (11), the E-step becomes:

$$
\begin{aligned}
Q\left(\theta|\theta^{(n)}\right) &= E_{Z|X,\theta^{(n)}}[\log L(\theta; X, Z)] \\
&= \sum_{g\in G} p(z_{g_1}, \ldots, z_{g_k}|X_g, \theta^{(n)}) \\
&\quad \times \left\{ \sum_{i\in M} \sum_{j\in c_g} \log \alpha_{i,j} + \log p(z_{g_1}, \ldots, z_{g_k}) \right\} \qquad (25)
\end{aligned}
$$

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

$$
\begin{aligned}
p(z_{g_1}, \ldots, z_{g_k}|X_g, \theta^{(n)}) &= \frac{p(z_{g_1}, \ldots, z_{g_k}; X_g, \theta^{(n)})}{p(X_g, \theta^{(n)})} \\
&= \frac{p(X_g, \theta^{(n)}|z_{g_1}, \ldots, z_{g_k}) p(z_{g_1}, \ldots, z_{g_k})}{\sum_{g_1,\ldots,g_k\in\mathcal{Y}_g} p(X_g, \theta^{(n)}|z_{g_1}, \ldots, z_{g_k}) p(z_{g_1}, \ldots, z_{g_k})} \\
&= \frac{\prod_{i\in M} \prod_{j\in c_g} \alpha_{i,j}\, p(z_{g_1}, \ldots, z_{g_k})}{\sum_{g_1,\ldots,g_k\in\mathcal{Y}_g} \prod_{i\in M} \prod_{j\in c_g} \alpha_{i,j}\, p(z_{g_1}, \ldots, z_{g_k})} \qquad (26)
\end{aligned}
$$

We note that $p(z_j = 1|X_j, \theta^{(n)})$ (i.e., $Z(n, j)$), 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 variable group $g$ that variable $C_j$ belongs to (i.e., $p(z_{g_1}, \ldots, z_{g_k}|X_g, \theta^{(n)}) \quad j \in c_g$). 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 $\theta^*$ that maximizes the $Q\left(\theta|\theta^{(n)}\right)$ function in each iteration to be the $\theta^{(n+1)}$ of the next iteration. Hence:

$$
\begin{aligned}
a_i^{(n+1)} &= a_i^* = \frac{\sum_{j\in SJ_i} Z(n, j)}{\sum_{j=1}^{N} Z(n, j)} \\
b_i^{(n+1)} &= b_i^* = \frac{\sum_{j\in SJ_i} (1 - Z(n, j))}{\sum_{j=1}^{N} (1 - Z(n, j))} \\
d_j^{t+1} &= d_j^* = Z(n, j) \qquad (27)
\end{aligned}
$$

where $Z(n, j) = p(z_j = 1|X_j, \theta^{(n)})$. 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:

$$
\begin{aligned}
p(z_{g_1}, \ldots, z_{g_k} | X_g, \theta^{(n)}) &= \frac{p(z_{g_1}, \ldots, z_{g_k}; X_g, \theta^{(n)})}{p(X_g, \theta^{(n)})} \\
&= \frac{p(X_g, \theta^{(n)} | z_{g_1}, \ldots, z_{g_k}) p(z_{g_1}, \ldots, z_{g_k})}{\sum_{g_1, \ldots, g_k \in \mathcal{Y}_g} p(X_g, \theta^{(n)} | z_{g_1}, \ldots, z_{g_k}) p(z_{g_1}, \ldots, z_{g_k})} \\
&= \frac{\prod_{i \in \mathcal{S}_j} \prod_{j \in c_g} \alpha_{i,j} p(z_{g_1}, \ldots, z_{g_k})}{\sum_{g_1, \ldots, g_k \in \mathcal{Y}_g} \prod_{i \in \mathcal{S}_j} \prod_{j \in c_g} \alpha_{i,j} p(z_{g_1}, \ldots, z_{g_k})} \\
&\quad \text{where } \mathcal{S}_j : \text{Set of sources observes } C_j
\end{aligned} \tag{28}
$$

$$
\begin{aligned}
a_i^{(n+1)} &= a_i^* = \frac{\sum_{j \in SJ_i} Z(n, j)}{\sum_{j \in \mathcal{C}_i} Z(n, j)} \\
b_i^{(n+1)} &= b_i^* = \frac{\sum_{j \in SJ_i} (1 - Z(n, j))}{\sum_{j \in \mathcal{C}_i} 1 - Z(n, j))} \\
d_j^{t+1} &= d_j^* = Z(n, j) \\
&\quad \text{where } \mathcal{C}_i \text{ is set of variables source } S_i \text{ observes}
\end{aligned} \tag{29}
$$

## References


Abdelzaher T et al (2007) Mobiscopes for human spaces. IEEE Pervasive Comput 6(2):20–29

Ali A, Khelil A, Szczytowski P, Suri N (2011) 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, ACM, New York, pp 67–76

Alur R, Courcoubetis C, Halbwachs N, Henzinger TA, Ho P-H, Nicollin X, Olivero A, Sifakis J, Yovine S (1995) The algorithmic analysis of hybrid systems. Theor Comput Sci 138(1):3–34

Burke J et al (2006) Participatory sensing. In: Workshop on world-sensor-web (WSW): mobile device centric sensor networks and applications, pp 117–134

Casella G, Berger R (2002) Statistical inference. Duxbury Press, Pacific Grove

Chang M, Ratinov L, Roth D (2012) Structured learning with constrained conditional models. Mach Learn 88(3):399–431 6

Cook B, Podelski A, Rybalchenko A (2005) Abstraction refinement for termination. In: Static analysis, Springer, pp 87–101

Cook B, Podelski A, Rybalchenko A (2006) Termination proofs for systems code. In: ACM SIGPLAN notices, ACM, vol 41. pp 415–426

Cramer H (1946) Mathematical methods of statistics. Princeton University Press, Princeton

de Valmaseda JM, Ionescu G, Deriaz M (2013) Trustpos model: trusting in mobile users location. In: Daniel F, Papadopoulos G, Thiran P (eds) Mobile web information systems, vol 8093., Lecture Notes in Computer ScienceSpringer, Berlin, pp 79–89

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. J R Stat Soc Ser B 39(1):1–38

Eisenman SB et al (2007) The bikenet mobile sensing system for cyclist experience mapping. In: SenSys'07

Ganti RK, Pham N, Ahmadi H, Nangia S, Abdelzaher TF (2010) Greengps: a participatory sensing fuel-efficient maps application. In: MobiSys '10: proceedings of the 8th international conference on mobile systems, applications, and services, ACM, New York, pp 151–164

Guitton A, Skordylis A, Trigoni N (2007) Utilizing correlations to compress time-series in traffic monitoring sensor networks. In: Wireless communications and networking conference, 2007. WCNC 2007. IEEE, pp 2479–2483

He L, Greenshields Ian R (2009) A nonlocal maximum likelihood estimation method for Rician noise reduction in MR images. Med Imaging IEEE Trans 28(2):165–172

He W, Liu X, Ren M (2011) Location cheating: a security challenge to location-based social network services. In: Distributed computing systems (ICDCS), 2011 31st international conference on, pp 740–749

Hoel PG et al (1954) Introduction to mathematical statistics, 2edn. John Wiley & Sons, Inc, New York

Hogg RV, Craig AT (1995) Introduction to mathematical statistics. Prentice Hall, New York

Hogg RV, McKean J, Craig AT (2005) Introduction to mathematical statistics. Prentice Hall, Inc

Hu S, Liu H, Su L, Wang H, Abdelzaher T (2013) SmartRoad: a mobile phone based crowd-sourced road sensing system. Technical report, University of Illinois at Urbana-Champaign, 08. https://www.ideals.illinois.edu/handle/2142/45699

Huang J-H, Amjad S, Mishra S (2005) CenWits: a sensor-based loosely coupled search and rescue system using witnesses. In: SenSys'05, pp 180–191

Hull B et al (2006) CarTel: a distributed mobile sensor computing system. In: SenSys'06, pp 125–138

Hunter T, Das T, Zaharia M, Abbeel P, Bayen AM (2012) Large scale estimation in cyberphysical systems using streaming data: a case study with smartphone traces. arXiv preprint. arXiv:1212.3393

Kay M, Choe EK, Shepherd J, Greenstein B, Watson N, Consolvo S, Kientz JA (2012) Lullaby: a capture & access system for understanding the sleep environment. In: Proceedings of the 2012 ACM conference on ubiquitous computing, ACM, pp 226–234

Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632

Koukoumidis E, Peh L-S, Martonosi M (2011) 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, ACM, New York, pp 353–354,

Lane ND, Miluzzo E, Eisenman SB, Musolesi M, Campbell AT (2008) Urban sensing systems: opportunistic or participatory

Lin T-H, Tarng W (1991) Scheduling periodic and aperiodic tasks in hard real-time computing systems. In: ACM SIGMETRICS performance evaluation review, vol 19. ACM, pp31–38

Liu CL, Layland JW (1973) Scheduling algorithms for multiprogramming in a hard-real-time environment. J ACM 20(1):46–61

Lu J, Sookoor T, Srinivasan V, Gao G, Holben B, Stankovic J, Field E, Whitehouse K (2010) The smart thermostat: using occupancy sensors to save energy in homes. In: Proceedings of the 8th ACM conference on embedded networked sensor systems, ACM, pp 211–224

Mok AK, Chen D (1997) A multiframe model for real-time tasks. Softw Eng IEEE Trans 23(10):635–645

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

Mun M, Reddy S, Shilton K, Yau N, Burke J, Estrin D, Hansen M, Howard E, West R, Boda P (2009) 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, ACM, New York, pp 55–68

Munir S, Stankovic JA, Liang C-JM, Lin S (2013) Cyber physical system challenges for human-in-the-loop control. In: Presented as part of the 8th international workshop on feedback computing. USENIX

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

Pandya M, Malek M (1998) Minimum achievable utilization for fault-tolerant processing of periodic tasks. Comput IEEE Trans 47(10):1102–1112

Park D-W, Natarajan S, Kanevsky A (1996) Fixed-priority scheduling of real-time systems using utilization bounds. J Syst Softw 33(1):57–63

Pasternack J, Roth D (2010) Knowing what to believe (when you already know something). In: International conference on computational linguistics (COLING)

Pham N, Ganti RK, Uddin YS, Nath S, Abdelzaher T (2010) Privacy-preserving reconstruction of multidimensional data maps in vehicular participatory sensing

Proietti T, Alessandra L (2012) Maximum likelihood estimation of time series models: the Kalman filter and beyond. MPRA paper, University Library of Munich, Munich

Qi G-J, Aggarwal CC, Han J, Huang T (2013) Mining collective intelligence in diverse groups. In: Proceedings of the 22nd international conference on world wide web, International World Wide Web Conferences Steering Committee, pp 1041–1052

Rachuri KK, Mascolo C, Musolesi M, Rentfrow PJ (2011) 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, ACM, New York, pp 73–84

Rajkumar RR, Lee I, Sha L, Stankovic J (2010) Cyber-physical systems: the next computing revolution. In: Proceedings of the 47th design automation conference, ACM, pp 731–736

Saeedloei N, Gupta G (2011) A logic-based modeling and verification of CPS. ACM SIGBED Rev 8(2):31–34

Scaglione A, Servetto SD (2002) 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, ACM, New York, pp 140–147

Schirner G, Erdogmus D, Chowdhury K, Padir T (2013) The future of human-in-the-loop cyber-physical systems. Computer 46(1):36–45

Sha L, Abdelzaher T, Årzén K-E, Cervin A, Baker T, Burns A, Buttazzo G, Caccamo M, Lehoczky J, Mok AK (2004) Real time scheduling theory: a historical perspective. Real-time Syst 28(2–3):101–155

Sha L, Gopalakrishnan S, Liu X, Wang Q (2009) Cyber-physical systems: a new frontier. In: Tsai JJP, Yu PS (eds) Machine learning in cyber trust. Springer, Berlin, pp 3–13

Sprunt B, Sha L, Lehoczky J (1989) Aperiodic task scheduling for hard-real-time systems. Real-Time Syst 1(1):27–60

Strosnider JK, Lehoczky JP, Sha L (1995) The deferrable server algorithm for enhanced aperiodic responsiveness in hard real-time environments. Comput IEEE Trans 44(1):73–91

Tang LA, Gu Q, Yu X, Han J, La Porta TF, Leung A, Abdelzaher TF, Kaplan LM (2012) Intrumine: mining intruders in untrustworthy data of cyber-physical systems. In: SDM, SIAM, pp 600–611

Uddin MYS, Wang H, Saremi F, Qi G-J, Abdelzaher T, Huang T (2011) Photonet: a similarity-aware picture delivery service for situation awareness. In: Proceedings of the 2011 IEEE 32nd real-time systems symposium, RTSS '11, IEEE Computer Society, Washington, DC, pp 317–326

Wang D, Abdelzaher T, Kaplan L, Aggarwal CC (2013) Recursive fact-finding: a streaming approach to truth estimation in crowdsourcing applications. In: The 33rd international conference on distributed computing systems (ICDCS'13)

Wang D, Abdelzaher T, Kaplan L (2015) Social sensing: building reliable systems on unreliable data. Morgan Kaufmann

Wang D, Abdelzaher T, Kaplan L, Ganti R, Hu S, Liu H (2013) Exploitation of physical constraints for reliable social sensing. In: The IEEE 34th real-time systems symposium (RTSS'13)

Wang D, Amin T, Li S, Abdelzaher T, Kaplan L, Gu S, Pan C, Liu H, Aggrawal C, Ganti R, Wang X, Mohapatra P, Szymanski B, Le H (2014) Humans as sensors: an estimation theoretic perspective. In: The 13th ACM/IEEE international conference on information processing in sensor networks (IPSN 14)

Wang D, Huang C (2015) Confidence-aware truth estimation in social sensing applications. In: The 12th annual IEEE international conference on sensing, communication, and networking

Wang D, Kaplan L, Abdelzaher T, Aggarwal CC (2012) On scalability and robustness limitations of real and asymptotic confidence bounds in social sensing. In: The 9th annual IEEE communications society conference on sensor, mesh and ad hoc communications and networks (SECON 12)

Wang D, Kaplan L, Le H, Abdelzaher T (2012) On truth discovery in social sensing: a maximum likelihood estimation approach. In: The 11th ACM/IEEE conference on information processing in sensor networks (IPSN 12)

Wang D, Kaplan LM, Abdelzaher TF, Aggarwal CC (2013) On credibility estimation tradeoffs in assured social sensing. IEEE J Sel Areas Commun 31(6):1026–1037

Wang D, Kaplan L, Abdelzaher TF (2014) Maximum likelihood analysis of conflicting observations in social sensing. ACM Transactions on Sensor Networks (TOSN) 10(2):30

Wang S, Wang D, Su L, Kaplan L, Abdelzaher TF (2014) Towards cyber-physical systems in social spaces: the data reliability challenge. In: Real-time systems symposium (RTSS), 2014 IEEE, IEEE, pp 74–85

Wolpaw J, Wolpaw EW (2012) Brain-computer interfaces: principles and practice. Oxford University Press, Oxford

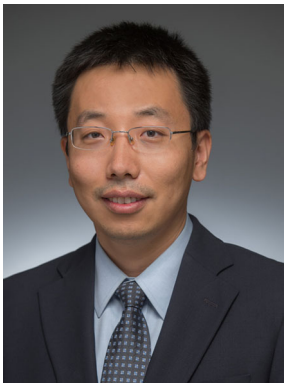
Wu CFJ (1983) On the convergence properties of the EM algorithm. Ann Stat 11(1):95–103

Xu Y, chien Lee W (2006) Exploring spatial correlation for link quality estimation in wireless sensor networks. In: Proceedings IEEE PerCom, pp 200–211

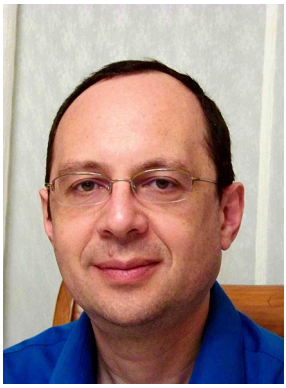
Yin X, Han J, Yu PS (2008) Truth discovery with multiple conflicting information providers on the web. IEEE Trans Knowl Data Eng 20:796–808

Zhao B, Rubinstein BIP, Gemmell J, Han J (2012) A bayesian approach to discovering truth from conflicting sources for data integration. Proc VLDB Endow 5(6):550–561

Zhou P, Zheng Y, Li M (2012) 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, ACM, New York, pp 379–392



**Dong Wang** received his Ph.D. in Computer Science from University of Illinois at Urbana Champaign (UIUC) in 2012. He is currently an assistant professor at the Department of Computer Science and Engineering, the University of Notre Dame. He has authored/co-authored more than 40 referred publications in the area of big data analytics, social sensing, cyber-physical computing, real-time and embedded systems. His interests lie broadly in developing analytic foundations for reliable information distillation systems, as well as the foundations of data credibility analysis, in the face of noise and conflicting observations, where evidence is collected by both humans and machines. He is the recipient of the Wing Kai Cheng Fellowship from University of Illinois in 2012 and the Best Paper Award of IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS) in 2010. He is a member of IEEE and ACM.



**Tarek Abdelzaher** received his Ph.D. from the University of Michigan in 1999 on Quality of Service Adaptation in Real-Time Systems. He is currently a Professor and Willett Faculty Scholar at the Department of Computer Science, the University of Illinois at Urbana Champaign. He has authored/coauthored more than 200 refereed publications in real-time computing, distributed systems, sensor networks, and control. He is an Editor-in-Chief of the Journal of Real-Time Systems, and has served as Associate Editor of the IEEE Transactions on Mobile Computing, IEEE Transactions on Parallel and Distributed Systems, IEEE Embedded Systems Letters, the ACM Transaction on Sensor Networks, and the Ad Hoc Networks Journal. He chaired (as Program or General Chair) several conferences in his area including RTAS, RTSS, IPSN, Sensys, DCoSS, ICDCS, and ICAC. Abdelzaher's research interests lie broadly in understanding and influencing performance and temporal properties of networked embedded, social and software systems in the face of increasing complexity, distribution, and degree of interaction with an external physical environment. Tarek Abdelzaher is a recipient of the IEEE Outstanding Technical Achievement and Leadership Award in Real-time Systems (2012), the Xerox Award for Faculty Research (2011), as well as several best paper awards. He is a member of IEEE and ACM.

**Lance Kaplan** received the B.S. degree with distinction from Duke University, Durham, NC, in 1989 and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1991and 1994, respectively, all in Electrical Engineering. From 1987–1990, Dr. Kaplan worked as a Technical Assistant at the Georgia Tech Research Institute. He held a National Science Foundation Graduate Fellowship and a USC Dean's Merit Fellowship from 1990?1993, and worked as a Research Assistant in the Signal and Image Processing Institute at the University of Southern California from 1993–1994. Then, he worked on staff in the Reconnaissance Systems Department of the Hughes Aircraft Company from 1994–1996. From 1996–2004, he was a member of the faculty in the Department of Engineering and a senior investigator in the Center of Theoretical Studies of Physical Systems (CTSPS) at Clark Atlanta University (CAU), Atlanta, GA. Currently, he is a researcher in the Networked Sensing and Fusion branch of the U.S Army Research Laboratory. Dr. Kaplan serves as Editor-In-Chief for the IEEE Transactions on Aerospace and Electronic Systems (AES) and as VP of Conferences for the International Society of Information Fusion (ISIF). Previously, he served on the Board of Governors of the IEEE AES Society (2008–2013) and on the Board of Directors of ISIF (2012–2014). He is a three time recipient of the Clark Atlanta University Electrical Engineering Instructional Excellence Award from 1999?2001. His current research interests include signal and image processing, information/data fusion, network science and resource management.

**Raghu Ganti** is a Research Staff Member at the IBM T. J. Watson Research center. He is part of the Cloud based networking department. His research interests span spatiotemporal analytics, big data, wireless sensor networks, privacy, and data mining. He obtained his MS and PhD degrees from the Department of Computer Science, University of Illinois, Urbana-Champaign in August 2010. He is the recipient of the Siebel scholar fellowship. He received his B.Tech degree from the Indian Institute of Technology, Madras in Computer Science and Engineering.

**Shaohan Hu** received his M.S. from Dartmouth College and is currently a Ph.D. candidate at the Department of Computer Science, the University of Illinois at Urbana-Champaign. His main research interests are in mobile sensing systems and sensor networks.

**Hengchang Liu** is currently an assistant professor at USTC. He got his Ph.D. degree at University of Virginia in 2011, under supervision of Professor John Stankovic. His research interest mainly includes cyber physical systems, mobile systems, named data networking, and wireless (sensor) networks.