Scalable Uncertainty-Aware Truth Discovery in Big Data Social Sensing Applications for Cyber-Physical Systems

Chao Huang, Dong Wang, and Nitesh V. Chawla

Abstract—Social sensing is a new big data application paradigm for Cyber-Physical Systems (CPS), where a group of individuals volunteer (or are recruited) to report measurements or observations about the physical world at scale. A fundamental challenge in social sensing applications lies in discovering the correctness of reported observations and reliability of data sources without prior knowledge on either of them. We refer to this problem as *truth discovery*. While prior studies have made progress on addressing this challenge, two important limitations exist: (i) current solutions did not fully explore the uncertainty aspect of human reported data, which leads to sub-optimal truth discovery results; (ii) current truth discovery solutions are mostly designed as sequential algorithms that do not scale well to large-scale social sensing events. In this paper, we develop a *Scalable Uncertainty-Aware Truth Discovery (SUTD)* scheme to address the above limitations. The SUTD scheme solves a constraint estimation problem to jointly estimate the correctness of reported data and the reliability of data sources while explicitly considering the uncertainty on the reported data. To address the scalability challenge, the SUTD is designed to run a Graphic Processing Unit (GPU) with thousands of cores, which is shown to run two to three orders of magnitude faster than the sequential truth discovery solutions. In evaluation, we compare our SUTD scheme to the state-of-the-art solutions using three real world datasets collected from Twitter: Paris Attack, Oregon Shooting, and Baltimore Riots, all in 2015. The evaluation results show that our new scheme significantly outperforms the baselines in terms of both truth discovery accuracy and execution time.

Index Terms—Big Data, Cyber-Physical Systems, Social Sensing, Uncertainty-Aware, Scalability, Truth Discovery, Parallel Implementation

1 Introduction

This paper presents a scalable uncertainty-aware estimation approach to solve the truth discovery problem in social sensing applications for Cyber-Phyiscal Systems (CPS). Social sensing has become a new big data application paradigm for CPS, where a group of individuals volunteer (or are recruited) to report measurements or observations about the physical world at scale [56]. Examples of social sensing applications include traffic monitoring and congestion control applications using data from drivers' or passengers' smartphones, geotagging and smart city applications using crowdsensing data from common citizens, and real-time situation awareness applications that report disaster fallout using online social media. Due to the open data contribution opportunities and unvetted nature of data sources (e.g., human sensors), a fundamental challenge in social sensing applications lies in *discovering the correctness of reported observations and reliability of data sources without prior knowledge on either of them*, which is referred to as *truth discovery problem* in social sensing. This work contributes to addressing the *veracity aspect* of the big data challenge in CPS applications.

Consider a disaster scenario like Ecuador Earthquake (April 2016), where many damages happened in the city and people volunteered to report real-time information about different aspects of the earthquake through online social media (e.g., Twitter). Such information can be effectively used to obtain accurate and timely situation awareness of the disaster and support decision makings on rescuing efforts and resource dispatch. However, it is challenging to accurately ascertain the correctness of human sensed data with little or no prior knowledge of the human sensors and the claims they contribute [66]. For example, users may report unreliable information on Twitter that could mislead people to the locations that do not have the desirable resources (e.g., food, water, gas) [17]. Furthermore, unlike physical sensors, humans are more likely to generate the claims with different degrees of uncertainty (e.g., affirmative assertions versus pure guesses), which add further complexity to the truth discovery problem [30].

Prior studies in sensor networks [36], [60], [63], data mining [12], [66], and machine learning communities [26], [40] have made a significant progress to address the truth discovery problem in social sensing. Despite such progress, two important limitations exist. First, current solutions did not fully explore the *uncertainty aspect* of the claims generated by human sensors and assumed all claims are *affirmative*. However, such assumption does not hold in real world social sensing applications. For example, during Oregon Shooting and Baltimore Riots events in 2015, people reported on Twitter their claims that are of different degrees of uncertainty in relation to the events (see Table 1). Simply ignoring such difference in uncertainty of claims are

- *The authors are with the Department of Computer Science and Engineering and the Interdisciplinary Center for Network Science and Applications (iCeNSA), University of Notre Dame, Notre Dame, IN 46556. E-mail: {chuang7,dwang5,nchawla}@nd.edu*

Events	Tweet	Uncertainty Degree
Oregon Shooting	"There's a shooter! Run! Run! Get out of there!" # OregonShooting	Low Uncertainty
Oregon Shooting	UNCOFIRMED: The Oregon school shooting may have been urged to action by 4chan members.	High Uncertainty
Baltimore Riots	5 things to know about Baltimore Mayor Stephanie Rawlings-Blake http://t.co/eniQSyXR9L	Low Uncertainty
Baltimore Riots	RT @JesusKreish: Baltimores throwin riots because this guy died?	High Uncertainty

Table 1: Claims of Different Degrees of Uncertainty in Real World Events

shown to lead to suboptimal truth discovery results [60], [63]. Second, current truth discovery solutions are mostly designed as *sequential algorithms* that cannot easily run on parallel computing platforms (e.g., cloud, GPU). Such scalability deficiency greatly limits the application of current truth discovery solutions in large-scale social sensing events.

A few technical challenges exist in order to address the above limitations of the truth discovery solutions. First, it is challenging to model and quantify the degrees of uncertainty human sensors express in their claims and incorporate such uncertainty feature into a rigorous truth discovery solution. Second, it is not a simple task to rigorously quantify the accuracy of the truth discovery results with the absence of the ground truth information on either source reliability or claim correctness. Third, it is nontrivial to design a parallel truth discovery solution that can run much faster than its sequential counterpart without sacrificing the truth discovery accuracy.

To address the above challenges, this paper develops a *Scalable Uncertainty-Aware Truth Discovery (SUTD)* scheme (Figure 1). The SUTD scheme solves a constraint estimation problem to jointly estimate the correctness of reported data and the reliability of data sources while explicitly exploring the uncertainty feature of claims. Rigorous confidence bounds have been derived to assess the quality of the truth discovery results output by SUTD scheme using the well-grounded results from estimation theory. We also designed a *parallel paradigm of SUTD* that runs a Graphic Processing Unit (GPU) with 2496 cores, which is shown to run two to three orders of magnitude faster than the sequential truth discovery solutions without degrading the performance in the estimation accuracy. In evaluation, we compare our SUTD scheme with state-of-the-art discovery solutions using three Twitter datasets collected during recent events: Paris Attack event, Oregon Shooting event and Baltimore Riots, all in 2015. The evaluation results demonstrate that our new scheme significantly improves both truth discovery accuracy and execution time compared to the baselines. In this paper, we primarily focus on the disaster and emergency response scenarios since the amount of factual and verifiable information is more significant compared to other social events (e.g., presidential election, protests). However, the authors discuss the limitation and possible generalization of our proposed model to better handle social events in Section 7. The results of this paper address two fundamental challenges in social sensing (i.e., uncertainty of claims and scalability of the solution), which provide a solid basis for future truth discovery solutions using principled approaches.

We summarized the contributions of this paper as follows:

- We explicitly address the uncertainty and scalability challenges of the truth discovery problem in social sensing. (Section 2)
- We developed a new analytical framework SUTD that solves the uncertainty-aware truth discovery problem using a principled approach in the context of big data social sensing applications. (Section 3)
- We implemented a parallel SUTD scheme on a GPU that was shown to run a few orders of magnitude faster than the sequential truth discovery solutions. (Section 4)
- We evaluated the performance of the SUTD scheme using three real world datasets collected from recent events. The evaluation results demonstrate the significant performance gain achieved by our scheme compared to other baselines. (Section 5)

An initial version of this work has been published in [61]. This work significantly expands on our previous work and makes new contributions from the following aspects. First, we extended our previous proposed model in [61] by developing new confidence bounds to rigorously assess the quality of the truth discovery results (Section 3). Second, we developed a scalable framework SUTD to implement our proposed scheme on a parallel platform (i.e., GPU), which can efficiently handle big data and is more suitable for large-scale social sensing events in big data applications (Section 4). Third, we compared our scheme with more recent truth discovery solutions from CPS literature and carried out a more comprehensive evaluation and comparison between the SUTD scheme and the state-of-the-art baselines (Section 5). Fourth, we performed a set of experiments on three new datasets collected from recent events (i.e., Paris attack, Oregon shooting and Baltimore riots in 2015) and further evaluated the robustness and efficiency of our scheme in these real world scenarios (Section 5). Finally, we extended our related work with specific discussion on Cyber-Physical Systems and discussed the fitness of our work into the scope of the special issue (Section 6).

2 Problem Statement and Distinction

2.1 Problem Statement

In this section, we introduce the uncertainty-aware truth discovery problem in social sensing as a constraint estimation problem. Specifically, let us consider a set of M sources, namely, S_1, S_2,...,S_M. The sources generate a collection of N claims about the physical world, namely, C_1, C_2,..,C_N. In this paper, we focus on binary claims. This is motivated by

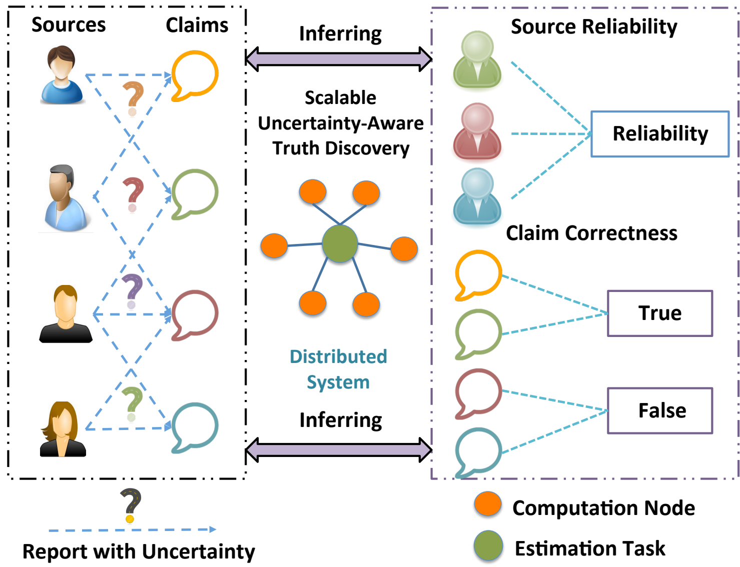


Figure 1: Overview of the SUTD Scheme

the observation that the states of the physical environment in many social sensing applications can be abstracted by a set of true or false statements. For example, in a geotagging application to find potholes on city streets, each possible location is associated with one claim that is true if a pothole presents at that location and false otherwise. Without loss of generality, we assume sources report only when a positive value is encountered (e.g., sources only report when she/he observes a pothole on streets). Let S_i represent the i^{th} source and C_j represent the j^{th} claim. $C_j = 1$ if it is true and $C_j = 0$ if it is false. Additionally, we introduce the definition of a *Sensing Matrix SC* where $S_iC_j = 1$ indicates that S_i reports C_j to be true, and $S_iC_j = 0$ otherwise.

In this paper, the uncertainty is defined as the degree of confidence (certainty) a source expresses in his/her report to a claim. In particular, we define an Uncertainty Matrix W, where the element $w_{i,j}$ denotes the degree of uncertainty source S_i expressed on claim C_j. We define the value of $w_{i,j}$ to be a discrete variable k, where $k \in [1, K]$ and K is the total number of degrees of uncertainty. In particular, $w_{ij} = k$ denotes that S_i reports the claim C_j to be true with a uncertainty degree of k, where $k = 1, ..., K$. The uncertainty degree k that a source expresses in its reports can be extracted from social sensing data using both syntactic (e.g., RT tag and URL of a tweet) and semantic features (uncertain words, replies from other users) of the claims. The details of the uncertainty degree computation are explained in Section 5.

In our model, we denote the reliability of source i as t_i. It indicates the likelihood that a claim is true if the source S_i reports it. In particular, t_i is given by:

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

Note that t_i is the overall reliability of a source S_i that incorporates all possible uncertainty degrees of S_i towards the claims he/she makes. It is not defined for a claim at a particular time instant.

Considering the fact that source S_i might have different reliability when it reports claims with different degrees of uncertainty [2], we define t_i^k as the reliability of source S_i when it reports a claim with an uncertainty degree of k (where $k = 1, ..., K$). Formally, t_i^k is given by:

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

Therefore,

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

where s_i^k is the probability that S_i reports C_j with a uncertainty degree of k. In particular, s_i^k can be computed based on the Uncertainty Matrix. Additionally, we denote the probability that S_i contributes a claim by s_i ($s_i = P(S_iC_j = 1)$). Note that $s_i = \Sigma_{k=1}^{K} s_i^k$.

We further define $T_{i,k}$ and $F_{i,k}$ to be the probability that source S_i reports claim C_j to be true with a uncertainty degree of k, given that the C_j is indeed true and false, respectively. In particular, $T_{i,k}$ and $F_{i,k}$ are defined as follows:

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

Following the Baye's theorem, relation between $T_{i,k}$, $F_{i,k}$ and t_i^k, s_i^k can be derived as follows:

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

where d is the probability that a randomly chosen claim is true, which is part of our estimation parameter defined in Section 3.

For completeness, we also define $T_i = P(S_iC_j = 1 | C_j = 1)$ and $F_i = P(S_iC_j = 1 | C_j = 0)$ to represent the overall probability that source S_i reports a claim to be true given the claim is true and false, respectively. The relationship between T_i, F_i and $T_{i,k}$, $F_{i,k}$ are as follows.

$$T_i = \Sigma_{k=1}^{K} T_i^k \times \frac{s_i^k}{s_i} \quad F_i = \Sigma_{k=1}^{K} F_i^k \times \frac{s_i^k}{s_i} \tag{6}$$

Therefore, the uncertainty-aware truth discovery problem in social sensing can be presented as a constraint estimation problem: given the input as the Sensing Matrix SC and Uncertainty Matrix W, the objective is to estimate the correctness of all claims and and the reliability of all sources. Formally, we compute:

$$\begin{aligned} &\forall j, 1 \leq j \leq N : P(C_j = 1|SC, W) \\ &\forall i, 1 \leq i \leq M : P(C_j = 1|S_iC_j = 1) \end{aligned} \tag{7}$$

2.2 Distinction from Previous Models

Before we present the SUTD scheme, we first highlight the difference between our model and a few closely related models from CPS and networked sensing literature [17], [36], [58], [60], [63].

Four recent models in truth discovery are most similar to our model: IPSN 12, RTSS 13, IPSN 14 and IPSN 16 model (shown in Figure 2). First, the IPSN 12 model is the seminal work that formulated the truth discovery problem as a network estimation problem [63]. Second, the RTSS 13 model extended the IPSN 12 model by considering the dependencies between claims. Both IPSN 14 and IPSN 16 considered the source dependence in the truth discovery problem. The difference between them are: the IPSN 14 simplified the source dependency graph as a set of two-level disjoint trees [60] while IPSN 16 developed a more generalized model to consider arbitrary source dependency graph (e.g., including multi-hop and cyclic dependency relationship) [17]. Moreover, the IPSN 16 also explicitly models the topic relevance feature of the claims. However, none of the above models studied the uncertainty aspect of the claims and the scalability of their schemes to large-scale social sensing events. In sharp contrast to previous work, this paper explicitly incorporates the *uncertainty on the reported data* and develops a *parallel truth discovery solution* to address the scalability problem. As shown in Figure 2, our model includes a set of variables to represent the uncertainty embedded in the claims and can run in parallel on a set of distributed nodes. The details of our SUTD schemed are presented in the following section.

3 An Uncertainty-Aware Truth Discovery (UTD) Scheme

In this section, we developed an UTD scheme using the Expectation-Maximization (EM) algorithm to solve the Uncertainty-Aware Truth Discovery problem. We also compute the Cramer-Rao Lower Bounds (CRLBs) to quantify the estimation accuracy of UTD scheme. In the next section, we extend the UTD scheme to SUTD scheme to address the scalablity challenge.

3.1 Background and Mathematical Formulation

We develop an uncertainty-aware Expectation Maximization (EM) to solve the constraint optimization problem formulated in the previous section. For the constraint estimation problem formulated in Section 2, the data we observed is Sensing Matrix SC and the Uncertainty Matrix W. The estimation parameter of our model is $\theta = (T_{1,k}, T_{2,k}, ..., T_{M,k}; F_{1,k}, F_{2,k}, ..., F_{M,k}; d)$ and $k = 1, 2, ..K$. $T_{i,k}$ and $F_{i,k}$ are defined in Equation (4) and d is the prior probability of a randomly chosen claim to be true. Furthermore, we introduce a vector of latent variables Z to represent the truthfulness of each claim. In particular, a variable z_j is defined for the j^{th} claim C_j: $z_j = 1$ if C_j is true and $z_j = 0$ otherwise. Additionally, in order to incorporate different degrees of uncertainty a source may express on her/his claims into the estimation problem, we define a set of binary variables w_{ij}^k such that $w_{ij}^k = 1$ if $w_{ij} = k$ in Uncertainty Matrix W and $w_{ij}^k = 0$ otherwise. The likelihood function of the uncertainty-aware truth discovery problem is as follows:

$$\begin{aligned} &L(\theta; X, Z) = \Pr(X, Z|\theta) \\ &= \prod_{j=1}^{Y} \Pr(z_j|X_j, \theta) \times \prod_{i=1}^{X} \prod_{k=1}^{K} \lambda_{i,j,k} \times \Pr(z_j) \end{aligned} \tag{8}$$

where $S_iC_j = 1$ when source S_i reports C_j to be true and 0 otherwise. Additional variables are defined in Table 2.

Table 2: Notations for UTD Scheme

$\lambda_{i,j,k}$	$\Pr(z_j)$	$Z(n,j)$	Constrains
$T_{i,k}$	d	$\Pr(Z_j = 1\|X_j, \theta^{(n)})$	$S_iC_j = 1, w_{i,j}^k = 1, z_j = 1$
$1 - \sum_{k=1}^{K} T_{i,k}$	d	$\Pr(Z_j = 1\|X_j, \theta^{(n)})$	$S_iC_j = 0, w_{i,j}^k = 1, z_j = 1$
$F_{x,k}$	$1-d$	$\Pr(Z_j = 0\|X_j, \theta^{(n)})$	$S_iC_j = 1, w_{i,j}^k = 1, z_j = 0$
$1 - \sum_{k=1}^{K} F_{i,k}$	$1-d$	$\Pr(Z_j = 0\|X_j, \theta^{(n)})$	$S_iC_j = 0, w_{i,j}^k = 1, z_j = 0$

3.2 UTD Scheme

Using the likelihood function, we derive the E-step as follows:

$$\begin{aligned} &Q(\theta|\theta^{(n)}) = R_{Z|X,\theta^{(n)}}[logL(\theta; X, Z)] \\ &= \sum_{j=1}^{N} Z(n,j) \times \sum_{i=1}^{M} (log\lambda_{i,j,k} + log\Pr(z_j)) \end{aligned} \tag{9}$$

We then define $Z(n, j) = p(z_j = 1|X_j, \theta^{(n)})$. It is the probability that a particular claim C_j is true given the observed data and current estimates of the parameters. $Z(n, j)$ can be further expanded as:

$$\begin{aligned} Z(n,j) &= \frac{p(z_j = 1; X_j, \theta^{(n)})}{p(X_j, \theta^{(n)})} \\ &= \frac{T(n,j) \times d^{(n)}}{T(n,j) \times d^{(n)} + F(n,j) \times (1 - d^{(n)})} \end{aligned} \tag{10}$$

where n is the iteration index. $T(n, j)$ and $F(n, j)$ are defined as follows:

$$\begin{aligned} &T(n,j) = p(X_j, \theta^{(n)}|z_j = 1) \\ &= \prod_{i=1}^{M} \prod_{k=1}^{K} (T_{i,k}^{(n)})^{S_iC_j \ \&\& \ w_{ij}^k} \times (1 - \sum_{k=1}^{K} (T_{i,k}^{(n)})^{1 - S_iC_j}) \\ &F(n,j) = p(X_j, \theta^{(n)}|z_j = 0) \\ &= \prod_{i=1}^{M} \prod_{k=1}^{K} (F_{i,k}^{(n)})^{S_iC_j \ \&\& \ w_{ij}^k} \times (1 - \sum_{k=1}^{K} (F_{i,k}^{(n)})^{1 - S_iC_j}) \end{aligned} \tag{11}$$

The Maximization step (M-step) is given by $\theta^{(n+1)} = \arg\max_{\theta} Q(\theta|\theta^{(n)})$. In the M-step, we select θ^* (i.e.,

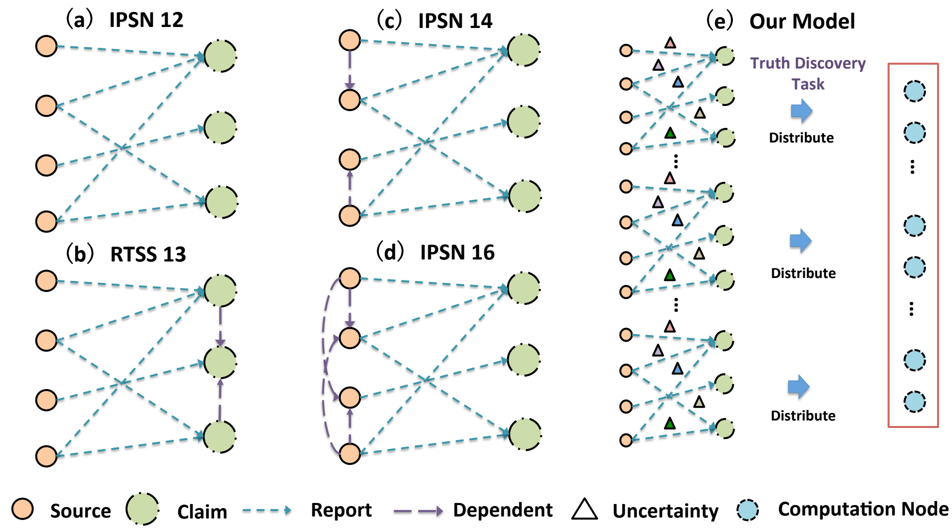


Figure 2: Comparison Between Our Model and Previous Models

$T_{1,k}, ..., T_{M,k}, F_{1,k}, ..., F_{M,k}, d$) that maximizes the $Q(\theta|\theta^{(n)})$ function in each iteration to be the $\theta^{(n+1)}$ of the next iteration.

To get θ^* that maximize $Q(\theta|\theta^{(n)})$, we solve $\frac{\partial Q}{\partial T_{i,k}} = 0$, $\frac{\partial Q}{\partial F_{i,k}} = 0$ and $\frac{\partial Q}{\partial d} = 0$ and get:

$$\sum_{j=1}^{N} \Big[Z(n,j) \times \Big((S_i C_j \; \&\& \; w_{ij}^k) \cdot \frac{1}{T_{i,k}^*} - (1 - S_i C_j) \frac{1}{1 - \Sigma_{h=1, h\neq k}^{K} T_{i,h}} \Big) \Big]$$
$$\sum_{j=1}^{N} \Big[Z(n,j) \times \Big((S_i C_j \; \&\& \; w_{ij}^k) \cdot \frac{1}{F_{i,k}^*} - (1 - S_i C_j) \frac{1}{1 - \Sigma_{h=1, h\neq k}^{K} F_{i,h}} \Big) \Big]$$
$$\sum_{j=1}^{N} \Big[Z(n,j) \cdot \frac{1}{d^*} - (1 - Z(n,j)) \cdot \frac{1}{1 - d^*} \Big] \tag{12}$$

Solving the above equations, we can obtain optimal $T_{i,k}^*$, $F_{i,k}^*$ and d^* are as follows:

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

where N represents the size of the claim set and SW_i^k represents the subset of claims that source S_i reports with the uncertainty degree of k. The UTD scheme is shown in Figure 3. Additionally, we summarize the UTD scheme in Algorithm 1.

3.3 Confidence Bounds of UTD Estimation

In this subsection, we present the derivation of the confidence bounds of the estimation results of the UTD scheme.

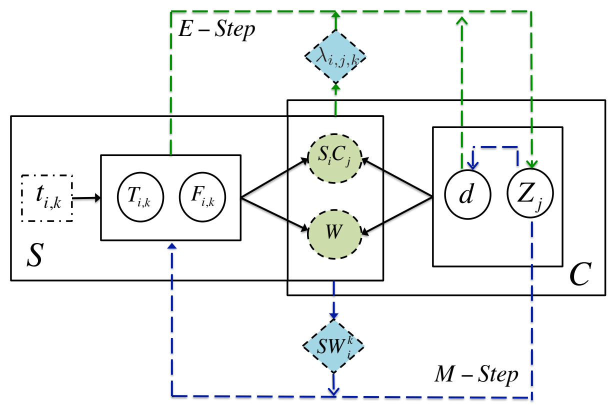


Figure 3: Probability Graphical Model of UTD

The confidences bounds are derived based on Cramer-Rao Lower Bounds (CRLB) of the estimations. Cramer-Rao Lower Bounds (CRLB) are the lowest bounds that can be reached by an unbiased estimator. It is defined as follows:

$$CRLB = J^{-1} \tag{14}$$

where J represents the Fisher information, which is a way of measuring the uncertainty (i.e., variance) on the estimation parameters of the model given the observed measurements [38].

Using the likelihood function defined in Equation (8), we can calculate the Fisher Information Matrix as follows:

$$(J(\hat{\theta}_{est}))_{i,j} = \begin{cases} 0 & i \neq j \\ -E_X[\frac{1}{\frac{\partial^2 l_{sutd}(x;T_{i,k})}{\partial T_{i,k}^2}}|_{T_{i,k}=\hat{T}_{i,k}^{est}}] & i = j \in [1, M] \\ -E_X[\frac{1}{\frac{\partial^2 l_{sutd}(x;F_{i,k})}{\partial F_{i,k}^2}}|_{F_{i,k}=\hat{F}_{i,k}^{est}}] & i = j \in (M, 2M] \end{cases} \tag{15}$$

Algorithm 1 UTD Algorithm

Initialize θ ($T_{i,k} = s_i^k, F_{i,k} = 0.5 \times s_i^k$, d =Random number in $(0, 1)$)
while $\theta^{(n)}$ does not converge **do**
  for $j = 1 : N$ **do**
    compute $Z(n, j)$ based on Equation (10)
  end for
  $\theta^{(n+1)} = \theta^{(n)}$
  for $i = 1 : M$ **do**
    compute $T_{i,k}^{(n+1)}, F_{i,k}^{(n+1)}, d^{(n+1)}$ based on Equation (13)
    update $T_{i,k}^{(n)}, F_{i,k}^{(n)}, d^{(n)}$ with $T_{i,k}^{(n+1)}, F_{i,k}^{(n+1)}, d^{(n+1)}$ in $\theta^{(n+1)}$
  end for
  $n = n + 1$
end while
Let Z_j^c = converged value of $Z(n, j)$
Let $T_{i,k}^c$ = converged value of $T_{i,k}^{(n)}$; $F_{i,k}^c$ = converged value of $F_{i,k}^{(n)}$; d^c = converged value of $d^{(n)}$
for $j = 1 : N$ **do**
  if $Z_j^c \geq threshold\ value$ **then**
    claim C_j is true
  else
    claim C_j is false
  end if
end for
for $i = 1 : M$ **do**
  calculate t_i^{k*} from $T_{i,k}^c$, $F_{i,k}^c$ and d^c based on Equation (4)
  calculate $t_i{}^*$ form t_i^{k*} based on Equation (3)
end for
Return the estimation on source reliability $t_i{}^*$ and corresponding judgment on the correctness of claim C_j.

where $\hat{T}_{i,k}^{est}$ and $\hat{F}_{i,k}^{est}$ are the converged values of estimation parameters derived in Equation (13). We can then obtain the CRLBs of our model by simply taking the inverse of the above Fisher information matrix.

Using the CRLB derived above, we can easily compute the derive confidence bounds of the estimation parameters [62]. In particular, the confidence bounds of T_i, F_i and t_i are computed as:

$$(\hat{t}_i^{est} - c_p\sqrt{var(\hat{t_i}^{est})}, \hat{t_i}^{est} + c_p\sqrt{var(\hat{t_i}^{est})}) \quad (16)$$

where $var(\hat{T_i}^{est})$ and $var(\hat{F_i}^{est})$ are the variance of the estimation parameters, which can be directly computed from the CRLBs in Equation (**??**). cp represents the standard score for confidence level p.

4 Scalable Uncertainty-Aware Truth Discovery (SUTD) Scheme

To address the scalability limitation of current truth discovery solutions, we develop a parallel implementation of the UTD scheme on a Graphic Processing Unit (GPU) using the Compute Unified Device Architecture (CUDA) programming model [35]. We refer to this parallel implementation of UTD as the *Scalable* Uncertainty-Aware Truth Discovery (SUTD) scheme. GPU has emerged as a new computing platform for many computational intensive applications. CUDA is a parallel programming model invented by NVIDIA. In CUDA, a *kernel* is defined as a grid of thread blocks and a thread of execution is the smallest unit in the parallelization. In the parallelization process, each node (called a *thread node*) will take care of a part of the whole computation task and users need to specify a set of *kernels* to parallelize the computation task.

Several challenges exist in order to implement SUTD: (i) the memory of Graphics Card is limited, so we need to design efficient strategies to handle the large-scale datasets on GPU; (ii) we need to design a mechanism to distribute the computation task of various estimation parameters and hidden variables of SUTD to different threads in an efficient way. To address these challenges, we designed the SUTD based on the estimation model developed in this paper and optimized our implementation using the following techniques: (i) we set the variables used in each thread as local variables instead of global variables given the fact that it costs more time to access global memory than local memory; (ii) we replaced the original conditional branch in the SUTD algorithm with a direct index in corresponding arrays, which allows us to save the waiting time of threads during the branch execution. The above optimization leads to significant execution time improvement achieved by SUTD as shown in the next section.

Algorithm 2 SUTD Algorithm

Input: Sensing Matrix Matrix SC, Uncertainty Matrix W
Output: Estimations of Source's Reliability and Claim's Correctness

Initialize θ ($T_{i,k} = r_i^k, F_{i,k} = 0.5 \times r_i^k$, d =Random number in $(0, 1)$)
$n = 0$
repeat
  $n = n + 1$
  CUDA Kernel of E-Step:
  for Each $j \in C$ **do**
    computation of $j \rightarrow$ one thread
    compute $\Pr(z_j = 1|X_j, \theta^{(n)})$
  end for
  CUDA Kernel of M-Step:
  for Each $i \in S$ **do**
    computation of $i \rightarrow$ one thread
    compute $(T_{i,k})^{(n)}, (F_{i,k})^{(n)}, (d)^{(n)}$
  end for
until $\theta^{(n)}$ and $\theta^{(n-1)}$ converge
The decision process is the same as the SUTD in Algorithm 1.

The main idea of SUTD is illustrated in Figure 4. Two key steps are designed to implement the SUTD: (1) we set up two different kernels, one for the E-step and the other for the M-step. (2) We allocate the computation tasks of E and M steps to different thread nodes. The independence of hidden variables and estimation parameters make the division of computational tasks and parallelization possible. Specifically, in the kernel of E-step, we distribute the computation task of hidden variables (i.e., Z_j) to N thread nodes. In the kernel of M-step, we distribute the computa-

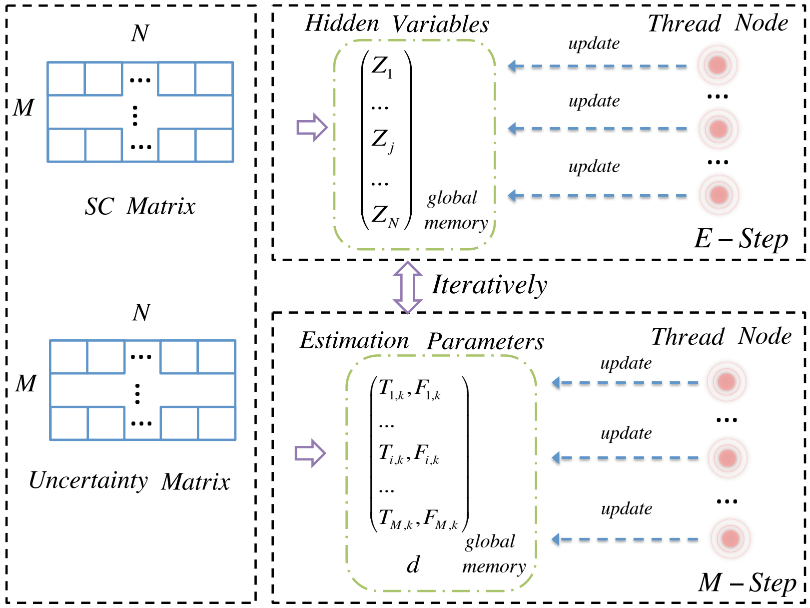


Figure 4: SUTD Scheme

tion task of estimation parameters (i.e., $T_{i,k}$, $F_{i,k}$ and d) to $2M \times K + 1$ thread nodes. We summarize the SUTD scheme in Algorithm 2.

5 Real World Case Studies

In this section, we evaluate the performance of the proposed SUTD scheme using three real-world data traces collected from Twitter. Twitter is an open data-sharing platform for average people and creates an ideal scenario for unreliable content from unvetted human sources with various degrees of uncertainty.

In our evaluation, we use the following truth discovery solutions as our baselines:

- *IPSN12* [63]: it solves the truth discovery problem using an iterative principle based on a maximum likelihood estimation framework.
- *IPSN14* [60]: it extended the IPSN 12 model by considering the dependencies between sources.
- *IPSN16* [17]: it extended the IPSN 14 model by considering the arbitrary source dependency and relevance of the claims to a given topic.
- *RTSS13* [58]: it addressed the truth discovery problem by exploring the dependency between claims.
- *HITS* [26]: it assumes that the relationship between source trustworthiness and claim's credibility is linear.
- *Majority Voting (MV)*: it simply assumes that a claim is more likely to be true if more sources report that claim.

Additionally, we also included the reference point called *Raw*, which represents the average percentage of true claims in a random sample set of raw tweets.

We have implemented the above schemes in the Apollo system, which is an information distillation framework the authors have developed to test truth discovery solutions in social sensing applications [60]. In particular, Apollo has two pre-processing components:

- *Data Collection Component*: it provides the interface for users to collect tweets using either keywords or geolocations based on the Twitter' search API.
- *Data Pre-processing Component*: it clusters tweets base on their contents using a variant of K-means clustering algorithm based on the Jaccard distance [47]. In particular, the Jaccard distance measures the overlap of keywords between any pair of compared tweets: the more overlapped keywords two tweets have, the shorter Jaccard distance they have.

We generated the the Sensing Matrix SC from the results of the data pre-processing steps. In particular, the sources are the Twitter users and the claims are the clusters of tweets that represent the collective observations from the social sensors.

The next step is to generate the Uncertainty Matrix W. In this paper, we focus on the binary case of claim uncertainty (i.e., $K = 2$). In particular, we use the following simple heuristics to roughly estimate the degree of uncertainty a user may express on a tweet. First, if the tweet is an original tweet (i.e., not a retweet) and contains a valid URL to an external source as the supporting evidence, it is of low uncertainty. Otherwise, it is of high uncertainty. The hypothesis of this heuristic is mainly twofold: (i) the first-hand information is often of lower uncertainty than the second-hand (e.g., retweet); (ii) including external evidence normally indicates stronger certainty of users. We call the first heuristic as *Syntactic* as it only uses the syntactic information of the tweets (e.g., RT tag or URL). Second, f the tweet does not contain any uncertain words and symbols (e.g., may, might and "?"), it is of low uncertainty. Otherwise, it is of high uncertainty. The hypothesis of this heuristic is that including uncertain words in the tweets normally indicates higher degree of uncertainty from users.

Data Trace	Paris Attack Event	Oregon Shooting Event	Baltimore Riots Event
Starting Date	11/13/2015	10/1/2015	4/14/2015
Duration of Trace	Eleven Days	Six Days	Seven Days
Physical Location	Paris, France	Umpqua, Oregon	Baltimore, Maryland
Search Keywords	Paris, Attacks, ISIS	Oregon, Shooting, Umpqua	Baltimore, Riots
Number of Tweets	873,760	210,028	952,442
Number of Users Tweeted	496,753	122,069	425,552

Table 3: Data Statistics of Three Traces

We refer to the second heuristic as *Semantic* as it considers the semantic information of tweets. Lastly, we consider the combination of the above two: if the tweet is an original tweet and contains a valid supporting URL as well as it does not contain any uncertain words, it is of low uncertainty. Otherwise, it is of high uncertainty. We refer to the third heuristic as *Syntactic+Semantic*. Note that the above heuristics are only approximations to estimate the degree of uncertainty a source may express on a tweet. In future, we will investigate deeper text analysis techniques (e.g., natural language processing) and study its impact on the claim uncertainty estimation.

In our evaluation, we select three real world Twitter data traces of recent events. The first trace collected tweets about Paris Attack in Nov, 2015. The second trace collected tweets about the Oregon Shooting that happened in Oct, 2015. The third one was collected from Baltimore Riots in April 2015. The reason for selecting those three data traces from disaster scenario is: those data traces contain more factual observations and their correctness can be verified from external resources. These traces are summarized in Table 3.

We fed each data trace to the Apollo system and executed all the compared truth discovery schemes. We manually graded the output of these schemes to determine the correctness of the claims. Considering the man-power limitations, we took the union of the top 50 claims returned by different schemes as our evaluation set. The following rubric is used to collect the ground truth labels for our evaluation:

- *True claims:* Claims that are statements of an event, which is generally observable by multiple independent sources and can be corroborated by credible sources external to Twitter (e.g., mainstream news media).
- *Undecided claims:* Claims that do not meet the criteria of true claims.

We note that undecided claims can potentially consist of two types of claims: (i) true claims that cannot be independently verified by external sources; (ii) false claims. Thus, our evaluation actually provides pessimistic performance bounds on estimations by treating undecided claims as false.

Also note that *SUTD scheme* is an parallel implementation of *UTD scheme*. We demonstrate in Section 4 that the parallelization implementation will not miss any information from the input data. In the following discussion, we just present the performance results of the *SUTD scheme*.

We first present performance results of SUTD scheme on Paris Attack data trace in Figure 5. The *SUTD-Syn*, *SUTD-Sem*, and *SUTD-Syn+Sem* represent the SUTD scheme that uses Syntactic heuristic, Semantic heuristic or both of them to infer the degree of uncertainty on claims. We observe that SUTD schemes generally outperform the compared baselines in most of the evaluation metrics: it discovers the most number of true claims while keeping the falsely reported one the least. Specifically, the largest performance gain is achieved by *SUTD-Syn*. The performance gain is 20% and 14% on accuracy and F1 score compared to the best performed baselines. The performance results on Oregon Shooting data trace are shown in Figure 6. The SUTD schemes continues to outperform all the baselines. The performance gain achieved by *SUTD-Syn+Sem* compared to the best performed baseline is 11% and 13% on accuracy and F1 score respectively. The results on Baltimore Riots data trace are shown in Figure 7. The results are consistent with previous experiments. The results on three real world data traces demonstrate that the SUTD schemes effectively identify truthful information in real-world applications where sources are unvetted and likely to express various degrees of uncertainty on their claims.

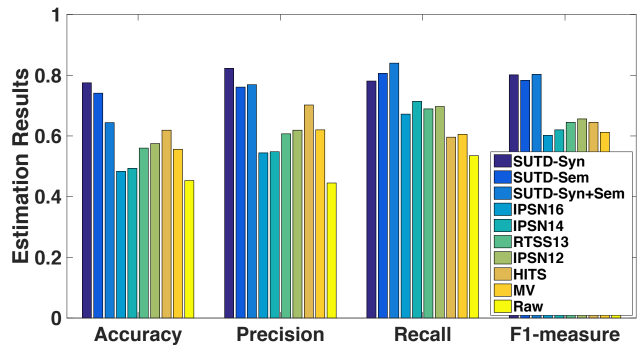


Figure 5: Evaluation on Paris Attack Trace

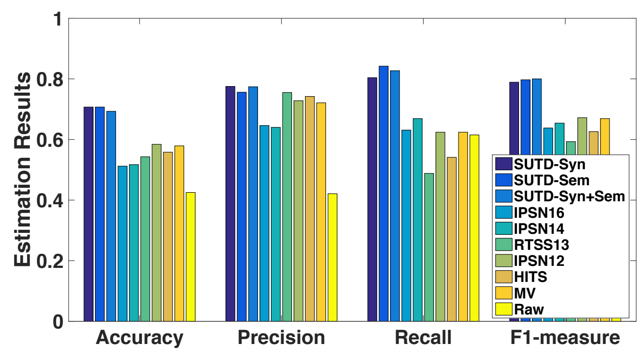


Figure 6: Evaluation on Oregon Shooting Trace

We would also like to understand whether the top truthful claims found by different algorithms actually capture the critical events that are newsworthy and reported by media. In particular, we independently collected 10 impor-

#	Media	Claim from SUTD	Claim from the Best Baseline
1	Gunman kills 9 at Oregon college, dies in shootout with police	Obama News Gunman kills nine at Oregon college, dies in shootout with police: By Courtney Sherwo... http://t.co/oOXSyh9OA0	**MISSING**
2	Oregon shooting: Gunman dead after college rampage	#cnn: Oregon shooting: Gunman dead at Umpqua Community College http://t.co/Ig3bbWYzFm #news	Oregon shooting: Gunman dead at Umpqua Community College #Umpqua #dead http://t.co/mf7D0dciEr
3	Witnesses Describe Chaotic Scene of Umpqua Community College Shooting	ABC Witnesses describe chaotic scene of shooting at Oregon college	RT @ABC: Witnesses describe chaotic scene of shooting at Oregon college: http://t.co/cdQhHgKlXO
4	Traumatized survivors tell of 'utter panic' during college shooting as heroic teachers evacuated students and others hid inside their classrooms	Umpqua Community College survivors tell stories from inside the shooting in Oregon: In the latest mass killing...	**MISSING**
5	Three pistols and a long rifle - were recovered from the scene.	RT @cnnbrk: Official: Three pistols and one rifle recovered at scene of Oregon shooting.	RT @cnnbrk: Official: Three pistols and one rifle recovered at scene of Oregon shooting.
6	10 killed, 7 injured at Oregon college shooting, officials say.	10 Killed in Shooting at Oregon Community College: Seven other people were injured and the gunman was ň... http://t.co/Wo1SJ6nl2W #news	RT @washingtonpost: Authorities confirm that 10 people were killed and seven others injured in the Oregon community college shooting
7	The gunman, identified by law enforcement as Chris Harper Mercer, 26, died in a gunfight with officers.	@CNN: Sources: Gunman at Oregon community college was 26-year-old Chris Harper Mercer.	**MISSING**
8	Oregon Sheriff Handling Massacre Fought the White House on Gun Control After Newtown	RT @YahooNews: Sheriff in #UCCShooting case fought the White House on gun control after Newtown massacre http://t.co/Y6Rrq8MaHm	RT @YahooNews: Sheriff in #UCCShooting case fought the White House on gun control after Newtown massacre http://t.co/Y6Rrq8MaHm
9	President Obama blames Congress for inaction on gun laws.	RT @nytimes: President Obama blames Congress for inaction on gun laws http://t.co/wn4ehMaMAU.	RT @nytimes: President Obama blames Congress for inaction on gun laws http://t.co/wn4ehMaMAU.
10	4chan thread under federal investigation after Oregon college shooting	4chan thread under federal investigation after Oregon college shooting http://t.co/ZfqLIGAOfz	4chan thread under federal investigation after Oregon college shooting http://t.co/1Rcgn8zOR1

Table 4: Ground truth events and related claims found by SUTD vs Best Performed Baselines in Oregon Shooting

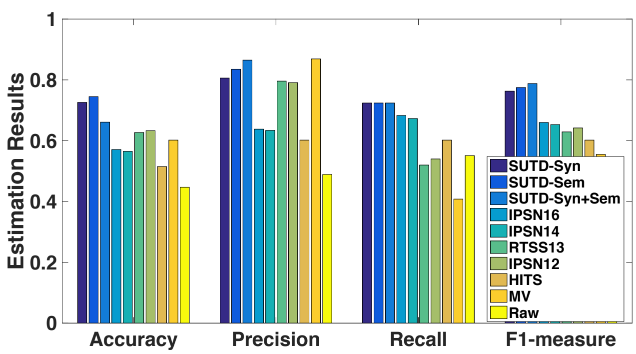


Figure 7: Evaluation on Baltimore Riots Trace

tant events covered by mainstream news media (e.g., CNN, BBC) during the Oregon Shooting event and used them as ground truths. After that we searched the top 50 ranked claims for each of the compared schemes to identify these events. We present the comparison results of the SUTD and the best performed baseline in Table 4. We observe that all ten milestone events are identified in the top claims returned by the SUTD scheme, while three of them are missing from the top claims returned from the best performed baseline. We repeated the same experiments on Paris Attacks and Baltimore Riots events and the results are similar: SUTD scheme found 8 milestone events in the case of Paris Attacks and 9 in Baltimore Riots compared to 6 and 7 by the best performed baseline.

Finally, we evaluate the efficiency of the parallel implementation of *SUTD scheme* discussed in Section 4. We implement SUTD on a computer with Nvidia GeFore GPU (2496 cores and 1.25 GHZ for each core, 4GB memory). We compare the SUTD with all baselines. We run the baselines on a regular lab computer (4 cores and 2 GHZ for each core, 8GB memory). Table 5 presents the execution time required by all algorithms on three data traces. We observe that the SUTD scheme runs several orders of magnitude faster than the compared baselines. The efficiency of SUTD is achieved by judiciously leveraging the computation power from thousands of cores on the GPU.

6 Related Work

Reliability is one of the fundamental challenges in Cyber-Physical Systems (CPS). Prior works in CPS have made significant advances to address the reliability challenge in time and functional dimensions [3], [9], [10], [28], [29], [34], [37], [39], [48], [51], [53], [65]. For the time reliability, there exist a rich amount of literature on designing various scheduling policies and utilization bounds in real time community [49]. For example, Liu et al. developed a set of basic utilization bounds for periodic tasks [29]. Many follow-up works extend the basic bounds by considering run-time [39], fault-tolerance [37], and multi-frame periodic models [34]. Utilization bounds have also been derived for aperiodic tasks [28], [51], [53]. For the functional reliability, it mainly focuses on correctness of program logic and system

Table 5: Execution Time Comparison (Seconds)

Algorithms	Paris Attack (s)	Oregon Shooting (s)	Baltimore Riots (s)
SUTD-Syn+Sem	**0.21**	**0.19**	**0.19**
SUTD-Syn	**0.25**	**0.19**	**0.19**
SUTD-Sem	**0.19**	**0.18**	**0.19**
IPSN16	572.42	437.38	217.10
IPSN14	620.15	400.34	221.23
RTSS13	71.06	54.76	47.42
IPSN12	63.83	42.73	51.49
HITS	13.81	9.41	10.01
MV	0.57	0.51	0.50

modeling [44], [50]. For example, Cook et al. developed useful tools for program analysis and software verification in cyber-physical and hybrid systems [9], [10]. Alur et al. and Saeeloei et al. developed formalism based methods to study the correctness of models in CPS [3], [48]. In contrast, this paper studies the *data reliability* challenge, which is motivated by the CPS applications with human-in-the-loop, especially the applications that use human as sensors.

Social sensing has emerged as a new application paradigm in CPS and smart cities [1], [2], [24], [31], [32], [33], [55]. The ideas of having humans involved in the process of sensing (e.g., participatory [6], [18], [22], opportunistic [27], [67] and human-centric [14], [20], [23] sensing) have been extensively studied in projects such as MetroSense [7], Urban Sensing [11] and SurroundSense [4]. The idea of using humans as sensors themselves came more recently [52]. For example, human sensors can contribute their observations through "sensing campaigns" [45], [46] or social data scavenging [21], [68]. A survey of social sensing [2] covers many challenges of using humans as sensors such as privacy perseverance [5], incentives design [25], and social interaction promotions [42], [43]. However, truth discovery remains to be a critical research question in social sensing. In this paper, we developed a new SUTD scheme to solve the uncertainty-aware truth discovery problem.

Fact-finders are a set of techniques developed in data mining and machine learning community to assess the quality of aggregated information from unreliable data sources. Hubs and Authorities [26] is one of the early fact-finders that computes the source and claim credibility in an iterative fashion. More fact-finding schemes have been developed to improve the basic frameworks by using probabilistic models [66], incorporating analysis on properties of claims [40] and dependency between sources [12]. More recent fact-finding algorithms address additional complexities such as prior knowledge on sources and claims [16], [19], [41] and the semantic features of claims [32], [59], [64]. In this paper, we will use insights from fact-finders and develop a new truth discovery solution that addresses uncertainty and scalability challenges in social sensing applications.

7 Discussions and Limitations

This paper presented a SUTD scheme that addressed two fundamental challenges in solving the truth discovery problem in social sensing: the uncertainty of reported data and the scalability of the solution. This work contributes to addressing the veracity aspect of the big data problem in CPS applications. While the current results are encouraging, there is room of further improvements. This section discusses some limitations we identified in the current SUTD scheme as well as the future work that we plan to carry out to address these limitations.

Sources are assumed to be independent in the current SUTD scheme. However, dependency may exist between sources, especially when they are connected through social networks. A set of social-aware truth discovery models have been recently developed to effectively address the source dependency problem in social sensing [17], [60]. On the other hand, no correlations are assumed between claims in our framework. The claim correlation problem has been studied by the authors in a separate line of work by incorporating the joint distribution on claim correlations into the truth discovery problem [58]. It worthy of noting that the aforementioned solutions on source dependency and claim correlation were developed under the same analytical framework as the SUTD scheme. This allows the authors to quickly develop a more generalized uncertainty-aware truth discovery model that explicitly considers both the source dependency and claim correlation under a unified framework.

The uncertainty estimation heuristics used in the SUTD scheme offer opportunities for future improvements. The Syntactic, Semantic, Syntactic+Semantic heuristics are only first approximations. Authors plan to improve them by leveraging more comprehensive techniques (e.g., text mining, natural language processing, etc.) to estimate the uncertainty of claims from a deeper analysis of the tweet contents. Some recent efforts provide good insights into this direction by developing new methods to exploit the lexicon, syntax and semantics of data from Twitter [13], [54]. Moreover, the uncertainty estimation module is a plug-in of the SUTD scheme, which gives us the flexibility to substitute it with a more refined one in the future.

In this paper, we mainly focused on the physical events (e.g., disaster and emergency scenarios). The reason is that the factual information is more significant in the physical events compared to the social events (e.g., president elections, protests) [56]. We also applied our model on the social events based datasets and the results are not very positive. The reasons are at least twofold: (i) There are a large amount of unfactual observations, sentiments and spams in the social events, which makes the truth discovery task in such context extremely challenging. (ii) The sources have a stronger social dependency in such events and misinformation and rumor spreading is much more significant compared to the physical events. We plan to further generalize the SUTD scheme to handle the unfactual claims and source dependency. In particular, we could model the factualness as an additional property of claims and integrate such property into the truth discovery framework. Moreover, we also plan to explicitly model the source dependency and incorporate such dependency into the SUTD scheme in a similar manner as other social-aware truth discovery framework [17], [60].

Considering the scope of the paper, we did not explicitly model the behavior of malicious users. Instead, we model the unknown source reliability in the SUTD scheme where the reliability of sources is not known to the social sensing applications *a priori*. Previous work have addressed malicious users detection problem and presented approaches to identify malicious users on social media [8], [15]. These results of the above work can be readily integrated with the SUTD scheme to solve the truth discovery with malicious users identification and removal as a pre-processing step. In particular, we will generalize the SUTD model by incorporating the malicious users detection results as prior knowledge, which will enforce a faster convergence of the EM algorithm and generate more accurate estimation results. Furthermore, we also plan to extend our current model to explicitly address source dependency and misinformation spread, which is critical to address the collusion attacks from the malicious users.

The time dimension of the problem deserves more investigation. When the uncertainty that a source expresses on claims changes with large dynamics over time, how to best account for it in the estimation framework? A time-sensitive model is needed to better handle such dynamics. Recent work in fact-finding literature starts to develop a new category of streaming EM algorithms that quickly update the estimation parameters using a recursive estimation approach [57]. Inspired by these results, the authors plan to develop similar real-time features of our SUTD scheme to better capture the dynamics in the uncertainty change. One key challenge is to design a nice tradeoff between estimation accuracy and computation complexity of the streaming algorithm. The authors are actively working on the above extensions.

8 Conclusion

This paper presents a Scalable Uncertainty-Aware Truth Discovery (SUTD) scheme to address two fundamental challenges that have not been well addressed in current truth discovery solutions: uncertainty of claims and scalability of algorithms. The SUTD scheme solves a constraint estimation problem to estimate both the correctness of reported data and the reliability of data sources while explicitly considering the uncertainty on the reported data. The SUTD scheme can run a Graphic Processing Unit with thousands of cores, which is shown to run a few orders of magnitude faster than current truth discovery solutions. We evaluated the performance of SUTD in comparison with the state-of-the-art baselines using three real world datasets collected from Twitter. The results show that SUTD scheme improves both the estimation accuracy and execution time of current truth discovery solutions. The results of this paper lay out a solid foundation to develop more scalalbe and accurate truth discovery models for big data social sensing applications in future research.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. CBET-1637251, CNS-1566465 and IIS-1447795 and Army Research Office under Grant W911NF-16-1-0388. 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 Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] T. Abdelzaher et al. Mobiscopes for human spaces. *IEEE Pervasive Computing*, 6(2):20–29, 2007.
[2] C. C. Aggarwal and T. Abdelzaher. Social sensing. In *Managing and Mining Sensor Data*, pages 237–297. Springer, 2013.
[3] R. Alur, C. Courcoubetis, N. Halbwachs, T. A. Henzinger, P.-H. Ho, X. Nicollin, A. Olivero, J. Sifakis, and S. Yovine. The algorithmic analysis of hybrid systems. *Theoretical computer science*, 138(1):3–34, 1995.
[4] M. Azizyan, I. Constandache, and R. Roy Choudhury. Surround-sense: mobile phone localization via ambience fingerprinting. In *Proceedings of the 15th annual international conference on Mobile computing and networking*, pages 261–272. ACM, 2009.
[5] I. Boutsis and V. Kalogeraki. Privacy preservation for participatory sensing data. In *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, volume 18, page 22, 2013.
[6] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava. Participatory sensing. *Center for Embedded Network Sensing*, 2006.
[7] A. T. Campbell, S. B. Eisenman, N. D. Lane, E. Miluzzo, and R. A. Peterson. People-centric urban sensing. In *Proceedings of the 2nd annual international workshop on Wireless internet*, WICON '06, New York, NY, USA, 2006. ACM.
[8] C. Cao and J. Caverlee. Detecting spam urls in social media via behavioral analysis. In *European Conference on Information Retrieval*, pages 703–714. Springer, 2015.
[9] B. Cook, A. Podelski, and A. Rybalchenko. Abstraction refinement for termination. In *Static Analysis*, pages 87–101. Springer, 2005.
[10] B. Cook, A. Podelski, and A. Rybalchenko. Termination proofs for systems code. In *ACM SIGPLAN Notices*, volume 41, pages 415–426. ACM, 2006.
[11] D. Cuff, M. Hansen, and J. Kang. Urban sensing: out of the woods. *Commun. ACM*, 51(3):24–33, Mar. 2008.
[12] X. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava. Global detection of complex copying relationships between sources. *PVLDB*, 3(1):1358–1369, 2010.
[13] M. Gupta, P. Zhao, and J. Han. Evaluating event credibility on twitter. In *SDM*, pages 153–164. SIAM, 2012.
[14] T. Higashino and A. Uchiyama. A study for human centric cyber physical system based sensing–toward safe and secure urban life–. In *Information Search, Integration and Personalization*, pages 61–70. Springer, 2013.
[15] X. Hu, J. Tang, and H. Liu. Online social spammer detection. In *AAAI*, pages 59–65, 2014.
[16] C. Huang and D. Wang. Spatial-temporal aware truth finding in big data social sensing applications. In *Trustcom/BigDataSE/ISPA, 2015 IEEE*, volume 2, pages 72–79. IEEE, 2015.
[17] C. Huang and D. Wang. Topic-aware social sensing with arbitrary source dependency graphs. In *2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 1–12. ACM/IEEE, 2016.
[18] C. Huang and D. Wang. Unsupervised interesting places discovery in location-based social sensing. In *Distributed Computing in Sensor Systems (DCOSS), 2016 International Conference on*, pages 67–74. IEEE, 2016.
[19] C. Huang, D. Wang, and N. Chawla. Towards time-sensitive truth discovery in social sensing applications. In *Mobile Ad Hoc and Sensor Systems (MASS)*, pages 154–162. IEEE, 2015.
[20] C. Huang, D. Wang, and B. Mann. Towards social-aware interesting place finding in social sensing applications. *Elsevier Knowledge Based Systems (KBS)*, 2017.

[21] C. Huang, D. Wang, and J. Tao. An unsupervised approach to inferring the localness of people using incomplete geo-temporal online check-in data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2017.

[22] C. Huang, D. Wang, and S. Zhu. Home location pro
ling of crowd sensors from noisy and sparse crowdsourcing data. In *Computer Communications, IEEE INFOCOM 2017-The 36th Annual IEEE International Conference on*. IEEE, 2017.

[23] C. Huang, D. Wang, S. Zhu, and D. Y. Zhang. Towards unsupervised home location inference from online social media. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 676–685. IEEE, 2016.

[24] C. Huang, X. Wu, and D. Wang. Crowdsourcing-based urban anomaly prediction system for smart cities. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1969–1972. ACM, 2016.

[25] L. G. Jaimes, I. J. Vergara-Laurens, and A. Raij. A survey of incentive techniques for mobile crowd sensing. *IEEE Internet of Things Journal*, 2(5):370–380, 2015.

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

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

[28] T.-H. Lin and W. Tarng. Scheduling periodic and aperiodic tasks in hard real-time computing systems. In *ACM SIGMETRICS Performance Evaluation Review*, volume 19, pages 31–38. ACM, 1991.

[29] C. L. Liu and J. W. Layland. Scheduling algorithms for multiprogramming in a hard-real-time environment. *Journal of the ACM (JACM)*, 20(1):46–61, 1973.

[30] M. Liu, S. Liu, X. Zhu, Q. Liao, F. Wei, and S. Pan. An uncertainty-aware approach for exploratory microblog retrieval. *IEEE transactions on visualization and computer graphics*, pages 250–259, 2016.

[31] J. Marshall, M. Syed, and D. Wang. Hardness-aware truth discovery in social sensing applications. In *12th IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS 16)*. IEEE, 2016.

[32] J. Marshall and D. Wang. Mood-sensitive truth discovery for reliable recommendation systems in social sensing. In *10th ACM Conference on Recommender Systems (Recsys 2016)*. ACM, 2016.

[33] J. Marshall and D. Wang. Towards emotional-aware truth discovery in social sensing applications. In *The 2nd IEEE International Conference on Smart Computing (SMARTCOMP 2016)*. IEEE, 2016.

[34] A. K. Mok and D. Chen. A multiframe model for real-time tasks. *Software Engineering, IEEE Transactions on*, 23(10):635–645, 1997.

[35] C. Nvidia. Programming guide, 2008.

[36] R. W. Ouyang, L. Kaplan, P. Martin, A. Toniolo, M. Srivastava, and T. J. Norman. Debiasing crowdsourced quantitative characteristics in local businesses and services. In *Proceedings of the 14th International Conference on Information Processing in Sensor Networks*, pages 190–201. ACM, 2015.

[37] M. Pandya and M. Malek. Minimum achievable utilization for fault-tolerant processing of periodic tasks. *Computers, IEEE Transactions on*, 47(10):1102–1112, 1998.

[38] V. Papathanasiou. Some characteristic properties of the fisher information matrix via cacoullos-type inequalities. *Journal of Multivariate analysis*, 44(2):256–265, 1993.

[39] D.-W. Park, S. Natarajan, and A. Kanevsky. Fixed-priority scheduling of real-time systems using utilization bounds. *Journal of Systems and Software*, 33(1):57–63, 1996.

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

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

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

[43] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas. Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 281–290. ACM, 2010.

[44] R. R. Rajkumar, I. Lee, L. Sha, and J. Stankovic. Cyber-physical systems: the next computing revolution. In *Proceedings of the 47th Design Automation Conference*, pages 731–736. ACM, 2010.

[45] S. Reddy, D. Estrin, and M. Srivastava. Recruitment framework for participatory sensing data collections. In *Proceedings of the 8th International Conference on Pervasive Computing*, pages 138–155. Springer Berlin Heidelberg, May 2010.

[46] S. Reddy, K. Shilton, G. Denisov, C. Cenizal, D. Estrin, and M. Srivastava. Biketastic: sensing and mapping for better biking. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 1817–1820, New York, NY, USA, 2010. ACM.

[47] K. D. Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking. Topical clustering of tweets. *Proceedings of the ACM SIGIR: SWSM*, 2011.

[48] N. Saeedloei and G. Gupta. A logic-based modeling and verification of cps. *ACM SIGBED Review*, 8(2):31–34, 2011.

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

[50] L. Sha, S. Gopalakrishnan, X. Liu, and Q. Wang. Cyber-physical systems: A new frontier. In *Machine Learning in Cyber Trust*, pages 3–13. Springer, 2009.

[51] B. Sprunt, L. Sha, and J. Lehoczky. Aperiodic task scheduling for hard-real-time systems. *Real-Time Systems*, 1(1):27–60, 1989.

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

[53] J. K. Strosnider, J. P. Lehoczky, and L. Sha. The deferrable server algorithm for enhanced aperiodic responsiveness in hard real-time environments. *Computers, IEEE Transactions on*, 44(1):73–91, 1995.

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

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

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

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

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

[59] D. Wang, M. T. Al Amin, T. Abdelzaher, D. Roth, C. R. Voss, L. M. Kaplan, S. Tratz, J. Laoudi, and D. Briesch. Provenance-assisted classification in social networks. *IEEE Journal of Selected Topics in Signal Processing*, 8(4):624–637, 2014.

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

[61] D. Wang and C. Huang. Confidence-aware truth estimation in social sensing applications. In *12th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 336–344. IEEE, 2015.

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

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

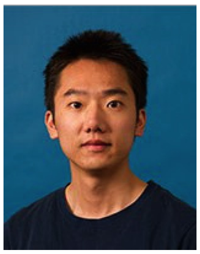
[64] D. Wang, J. Marshall, and C. Huang. Theme-relevant truth discovery on twitter: An estimation theoretic approach. In *Tenth International AAAI Conference on Web and Social Media (ICWSM)*, 2016.

[65] J. Wang, D. Wang, Y. Zhao, and T. Korhonen. Fast anti-collision algorithms in rfid systems. In *Mobile Ubiquitous Computing, Systems, Services and Technologies, 2007. UBICOMM'07. International Conference on*, pages 75–80. IEEE, 2007.

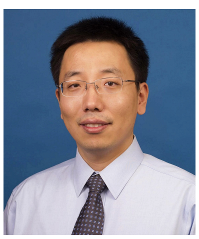
[66] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE Trans. on Knowl. and Data Eng.*, 20:796–808, June 2008.

[67] D. Y. Zhang, R. Han, D. Wang, and C. Huang. On robust truth discovery in sparse social media sensing. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 1076–1081. IEEE, 2016.

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



Chao Huang is a PhD student in the Department of Computer Science and Engineering at the University of Notre Dame, and is affiliated with the University's Interdisciplinary Center for Network Science and Applications (iCeNSA). His research focuses on big data analytics, with an emphasis on social sensing and social network analysis.



Dong Wang received his Ph.D. in Computer Science from University of Illinois at Urbana Champaign (UIUC) in 2012, an M.S. degree from Peking University in 2007 and a B.Eng. from the University of Electronic Science and Technology of China in 2004, respectively. He is now an assistant professor in the Department of Computer Science and Engineering at University of Notre Dame. He is also affiliated with the Interdisciplinary Center for Network Science and Applications (iCeNSA). Dr. Wang's research interests lie in the area of big data analytics, social sensing, cyber-physical computing, smart city, and crowdsourcing applications. He received 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.

Nitesh Chawla is the Frank M. Freimann Professor of Computer Science and Engineering at the University of Notre Dame. He is also the director in the Interdisciplinary Center for Network Science and Applications (iCeNSA), an institute focused on network and data science. His research work has led to many interdisciplinary contributions in social networks, healthcare analytics, environmental sciences, learning analytics, and media. He has received prestigious recognitions such as the IBM Big Analytics Award, IBM Watson Faculty Award, IEEE CIS Outstanding Early Career Award, National Academy of Engineers New Faculty Fellowship, and Outstanding Teacher Awards.