Topic-Aware Social Sensing with Arbitrary Source Dependency Graphs

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

***Abstract*—This work is motivated by the emergence of social sensing as a new paradigm of collecting observations about the physical environment from humans or devices on their behalf. These observations may be true or false, and hence are viewed as binary claims. A fundamental problem in social sensing applications lies in ascertaining the correctness of claims and the reliability of data sources without knowing either of them *a priori*. We refer to this problem as *truth discovery*. Prior works have made significant progress to addressing the truth discovery problem, but two significant limitations exist: (i) they ignored the fact that claims reported in social sensing applications can be either relevant or irrelevant to the topic of interests. (ii) They either assumed the data sources to be independent or the source dependency graphs can be represented as a set of disjoint trees. These limitations led to suboptimal truth discovery results. In contrast, this paper presents the first social sensing framework that explicitly incorporates the topic relevance feature of claims and arbitrary source dependency graphs into the solutions of truth discovery problem. The new framework solves a multi-dimensional maximum likelihood estimation problem to jointly estimate the truthfulness and topic relevance of claims as well as the reliability and topic awareness of sources. We compared our new scheme with the state-of-the-art truth discovery solutions using three real world data traces collected from Twitter in the aftermath of Paris Shooting event (2015), Hurricane Arthur (2014) and Boston Bombing event (2013) respectively. The evaluation results showed that our schemes significantly outperform the compared baselines by identifying more relevant and truthful claims in the truth discovery results.**

***Index Terms*—Social Sensing, Truth Discovery, Topic-Aware, Arbitrary Source Dependency Graph, Twitter**

I. Introduction

This paper presents a principled framework to address the topic-aware truth discovery problem with arbitrary source dependency graphs in social sensing applications. Social sensing has emerged as a new paradigm of collecting observations about the physical environment from humans or devices on their behalf [2]. This paradigm is motivated by the proliferation of various sensors in the possession of common individuals and the popularity of social networks that enable massive information dissemination opportunities. For example, drivers may contribute data through their smartphones to report the state of traffic congestion at various locales [7]. Alternatively, survivors may contribute data to online social media (e.g., Twitter, Facebook, Flickr) to document the damage and outage in the aftermath of a disaster [24]. These observations may be true or false, and hence are viewed as binary *claims*. A fundamental problem in social sensing applications lies in accurately ascertaining the truthfulness of claims and the reliability of data sources. We refer to this problem as *truth discovery*.

Consider an emergency response scenario (e.g., campus shooting, fire disaster, or bombing event) as an example. Survivors and witnesses may spontaneously report to online social media (e.g., Twitter, Facebook, Google+) about the damage and the current situation of the event (e.g., the shooter's location, available exit route, number of victims). Some reports are true and some are false. Without knowing the individual sources *a priori*, it is very challenging to identify the truthfulness of each claim. Majority voting (i.e., simply counting the number of sources that ascertain the same claim) is not always a good measure of claim truthfulness, as different sources may have different reliability [33]. In fact, the reliability of individual sources is not known in advance in such applications (i.e., we normally do not know when and where the emergent event will happen and who will get involved) [27]. Moreover, sources may use the keyword of the emergent event (e.g., hashtag in Twitter) to generate irrelevant claims with a purpose of attracting more public attention [4]. Additionally, sources could also intentionally or unintentionally forward misinformation through their social networks [30]. All these complexities make the truth discovery in social sensing a non-trivial task.

Previous works have made significant progress to address the truth discovery problem in social sensing [10], [11], [17], [30], [31], [33], [37], [40]. However, two significant limitations exist in the state-of-the-arts solutions. First, current solutions ignored the fact that claims reported in social sensing

Tweet	Topic Relevance
Thanks to generosity of volunteer blood donors there is currently enough blood on the shelves to meet demand. #BostonMarathon	Relevant
Child killed in Monday's Boston Marathon bombings identified as 8-year-old Martin Richard, reports the Boston Globe	Relevant
Over at Canary Whalf! Sun is shinning! Making me excited for the weekend antics! #Party #BostonMarathon#Bbq	Irrelevant
Next month, a special ONE DAY ONLY offer is coming #half #BostonMarathon	Irrelevant

Table I
On and Off Topic Claims in Boston Marathon Bombing Event

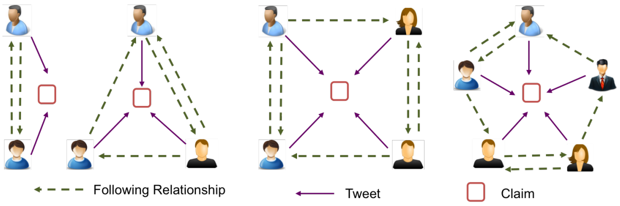


Figure 1. Source Cyclic Dependency Examples in Twitter

applications can be either relevant or irrelevant to the topic of interests [37]. For example, in the aftermath of the Boston Marathon Bombing event in 2013, people reported their claims (e.g., tweets) to online social media that are both relevant and irrelevant to the topic of the bombing event (Table I). It is very difficult (if possible) to find a set of keywords that could clearly separates all relevant claims from the irrelevant ones, especially with no prior knowledge of a particular event. Running truth discovery schemes on all collected claims without considering whether a claim is relevant to the topic of interests or not will generate many irrelevant claims that could significantly interfere with decision-making [17]. Second, data sources in social sensing are humans or devices they operate. It is not unusual for a human source to report claims they received from others. The previous truth discovery techniques either completely ignored the source dependency problem [33] or assumed source dependencies can be represented by graphs of disjoint trees [30]. This oversimplification on the source dependency leads to suboptimal truth discovery results since many dependency links are ignored in this abstraction (e.g., DAG dependency, cyclic dependency). Figure 1 shows some simple examples extracted from Twitter where sources with cyclic dependencies report the same claim. Such cyclic dependencies make it very challenging to accurately identify the provenance of reported claims (e.g., who reports the original claim). The uncertain data provenance was shown to have a direct impact on the truth discovery results (e.g., a group of dependent sources could generate a misinformed claim by simply repeating it many times on social media, which might confuse the truth discovery solutions) [33]. This problem has not been fully addressed in the previous study [30] where source dependencies were assumed to be *cycle free*. This paper develops a new model to explicitly handle arbitrary source dependency graphs.

Important challenges exist when we generalize the truth discovery problem in social sensing by explicitly considering the topic relevance of claims and arbitrary source dependency graphs. First, social sensing is designed as an open data collection paradigm where the reliability (the likelihood of a source to report truthful claims) and the topic awareness (the likelihood of a source to report topic relevant claims) of sources are often *unknown* a priori. Second, filtering out the irrelevant claims by simply using predefined keywords (or the hashtags in the Twitter examples) is not sufficient or possible due to several reasons: (i) some relevant tweets may not contain the predefined keywords (e.g., people can use various words to describe the same event); (ii) some claims that contain the keywords related with the topic of interests are actually irrelevant (e.g., in order to attract attention). Third, the source dependency graphs can be arbitrary since any pair of data sources could potentially be related in social sensing. It is a significant challenge to generalize the social sensing framework to handle arbitrary source dependency graphs.

In this paper, we present the first analytical framework that explicitly incorporates the *topic relevance* feature of claims and *arbitrary source dependency graphs* into the solutions of truth discovery problem in social sensing. The new framework solves a multi-dimensional maximum likelihood estimation problem where the topic relevance feature of claims is modeled as a vector of hidden variables and the arbitrary source dependencies are encoded into the estimation parameters. In particular, two new Expectation Maximization (EM) based algorithms have been developed: Topic-Aware EM (TA-EM) and Topic-Aware Source-Dependent EM (TASD-EM). These new algorithms jointly assign true and topic relevance values to claims and reliability and topic awareness values to sources in a way that is most consistent with the observed social sensing data. We compared our new schemes with the state-of-the-art truth discovery solutions using three real world data traces collected from Twitter in the aftermath of Paris Shooting event in 2015, Hurricane Arthur in 2014, and Boston Bombing event in 2013 respectively. The evaluation results showed that our schemes significantly outperform the compared baselines by identifying more relevant and truthful claims in the truth discovery results. The results of this paper are important because they allow social sensing applications to accurately estimate the truthfulness and relevance of claims as well as the reliability and topic awareness of sources by excluding irrelevant and false claims from the final estimation results using a principled approach.

In summary, our contributions are as follows:

- To the best of our knowledge, this study is the first to explicitly consider both the topic relevance feature of claims and arbitrary source dependency graphs in solving the truth discovery problem in social sensing.
- We develop a principled framework that allows us to derive optimal solutions (in the sense of maximum likelihood estimation) that are most consistent with the observed social sensing data and source dependencies.
- We perform extensive experiments to investigate the performance of our schemes and other truth discovery solutions on real world social sensing data traces. The evaluation results demonstrate the effectiveness and non-trivial performance gains achieved by our new schemes.

The rest of this paper is organized as follows: we discuss the related work in Section II. In Section III, we present the new topic-aware truth discovery model with arbitrary source dependency graphs. We compare our model with previous truth discovery models in Section IV. The proposed TA-EM and TASD-EM algorithm are presented in Section V and Section VI respectively. We present the experiments and evaluation results in Section VII. The limitations and future work are discussed in Section VIII. Finally, we conclude the paper in Section IX.

II. Related Work

Social sensing has emerged as a new sensing paradigm that empowers average people to contribute their observations and measurements about the physical world at a very large scale [1]. A comprehensive overview of social sensing applications is presented in [26]. More recent works have focused on addressing important challenges such as mobile audio sensing [34], data fusion and information aggregation [29], [36], distributed cloud sensing [42], and social group stability [19]. Truth discovery is a critical problem for reliable social sensing applications [2]. Previous work made important progress to address this problem but their solutions either ignored the topic relevance feature of claims [30], [33] or made oversimplified assumption on source dependencies [17], [37]. In contrast, this paper develops a new analytical framework that provides a generalized truth discovery solution by explicitly considering both the *topic relevance* feature of claims and *arbitrary source dependencies* in social sensing applications.

In data mining and machine learning literature, there exists a good amount of work on the topics of *fact-finding* that jointly compute the source reliability and claim credibility [8]. *Hubs and Authorities* [15] established a basic fact-finding model based on linear assumptions to compute scores for sources and claims they asserted. Yin et al. introduced *TruthFinder* as an unsupervised fact-finder for trust analysis on a providers-facts network [39]. Other fact-finders enhanced these basic frameworks by incorporating analysis on properties or dependencies within claims and sources [21], [25]. More recently, new fact-finding algorithms have been designed to address the background knowledge [18], multi-valued facts [43], and multi-dimensional aspects of the problem [41]. Using the insights (i.e., the mutual dependency between source reliability and claim correctness) from the above work, we develop a new estimation framework to explicitly model unreliable human sensors and solve the topic aware truth discovery problem in social sensing applications.

Our work is also related with reputation and trust systems that are designed to assess the reliability of sources (e.g., the quality of providers) [3], [38]. eBay is a typical reputation system based on a homogeneous peer-to-peer network structure, which allows participants to rate each other after each pair of them conduct a transaction [9]. Alternatively, Amazon on-line review system represents another type of reputation system based on a heterogeneous network structure, where different sources offer reviews on products (or brands, companies) they experienced [6]. Recent work has also investigated the consistency of reports to estimate and revise trust scores in reputation systems [12]–[14]. However, in social sensing, we normally do not have enough history data to compute the converged reputation scores of sources due to the short-lived sensing campaigns [27]. Instead, this paper presents a maximum likelihood estimation approach that jointly estimates the reliability and topic awareness of sources as well as the truthfulness and topic relevance of claims based on the observations collected from social sensing applications.

Finally, maximum likelihood estimation (MLE) framework has been widely used in the wireless sensor network (WSN) and data fusion communities [16], [20], [23], [35]. For example, Pereira et al. proposed a diffusion-based MLE algorithm for distributed estimation in WSN in the presence of data faults [20]. Sheng et al. developed a MLE method to infer locations of multiple sources by using acoustic signal energy measurements [23]. Eric et al. deisgned a MLE based approach to aggregate the signals from remote sensor nodes to a fusion center without any inter-sensor collaborations [16]. However, the above work primarily focused on the estimation of continuous variables from physical sensor measurements. In contrast, this paper focuses on a set of *binary variables* that represent either true/false and relevant/irrelevant claims from human sensors. The discrete nature of the estimation variables leads to a more challenging optimization problem that has been solved in this paper.

III. Problem Formulation

In this section, we formulate our topic-aware truth discovery problem with arbitrary source dependency graphs as a multi-dimensional maximum likelihood estimation problem. In particular, we consider a social sensing scenario where a group of M sources $S = (S_1, S_2, ..., S_M)$ report a set of N claims $C = C_1, C_2, ..., C_N$. In this paper, we consider two independent features of a claim: (i) topic relevance: whether a claim is related to the topic of interests or not; (ii) truthfulness: whether a claim is true or false. We let S_u denote the u^{th} source and C_k denote the k^{th} claim. $C_k = O$ and $C_k = \overline{O}$ represent that claim C_k is relevant or irrelevant to the topic of interests respectively. In social sensing applications, sources may indicate a claim to be relevant to a certain topic (e.g., using hashtags in Twitter). Furthermore, $C_k = T$ and $C_k = F$ represent the claim to be true or false respectively. We further define the following terms to be used in our model.

- ST is defined as a $M \times N$ matrix to represent whether a source indicates a claim to be topic relevant or not. It is referred to as the *Source-Topic Matrix*. In ST, $S_uT_k = 1$ when source S_u indicates C_k to be relevant to a topic of interests and $S_uT_k = -1$ when source S_u does not indicate C_k to be topic relevant and $S_uT_k = 0$ if S_u does not report C_k at all.
- SC is defined as a $M \times N$ matrix to represent whether a source reports a claim to be true. It is referred to as the *Source-Claim Matrix*. In SC, $S_uC_k = 1$ if source S_u reports the claim C_k and $S_uC_k = 0$ otherwise. Following the previous social sensing models [26], we assume that a source will only report the positive status of a claim (e.g., in a social sensing application to report potholes on city streets, sources will only generate claims when they observe potholes).
- SD is defined as a $M \times M$ matrix that represents source dependencies. It is referred to as the *Source-Dependency Matrix*. In SD, $SD_{u,v} = 1$ if source S_u and source S_v

have a directed dependency connection (e.g., S_v retweets S_u or replies to S_u in Twitter) and $SD_{u,v} = 0$ otherwise. Based on the SD matrix, we can partition the whole set of sources into C independent groups where sources in different independent groups have zero elements in SD.

Note that the derivations of SC, ST and SD are explained in the evaluation (i.e., Section VII).

One key challenge in social sensing applications lies in the fact that sources are often unvetted and they may not always report relevant and truthful claims. Hence, we need to explicitly model both the topic awareness and reliability of sources. First, we define the *topic-awareness* of source S_u as Ta_u: the probability that a claim C_k is topic relevant given the source S_u indicates it to be. Second, we observe that the source reliability is directly related with the source dependencies (e.g., an independent report should be treated differently from a repeated report in the calculation of source reliability) [30]. Hence, we define the reliability for both independent and dependent sources. If source S_u is an independent source, we define the *independent reliability* of source S_u as Re_u: the probability that a claim is true given that source S_u reports it to be true. If source S_u is dependent, we define $Re_{u,v}$ as the source's *dependent reliability*: the probability that source S_v (a dependent source of S_u) reports a claim to be true and the claim is indeed true given that S_u reports it to be true. Formally, Ta_u, Re_u and $Re_{u,v}$ are defined as follows:

$$\begin{aligned}
Ta_u &= \Pr(C_k = O | S_u T_k = 1)\\
Re_u &= \Pr(C_k = T | S_u C_k = 1)\\
Re_{u,v} &= \Pr(C_k = T, S_v C_k = 1 | S_u C_k = 1)
\end{aligned} \tag{1}$$

We further define a few conditional probabilities that we will use in our problem formulation. Specifically, we define $E_{u,O}^T$ and $E_{u,O}^F$ as the (unknown) probability that source S_i reports a claim to be topic relevant or not given the claim is indeed topic relevant. Similarly, we define $E_{u,\overline{O}}^T$ and $E_{u,\overline{O}}^F$ as the (unknown) probability that source S_i reports a claim to be topic relevant or not given the claim is indeed topic irrelevant. Formally, $E_{u,O}^T$, $E_{u,O}^F$, $E_{u,\overline{O}}^T$ and $E_{u,\overline{O}}^F$ are defined as:

$$\begin{aligned}
E_{u,O}^T &= \Pr(S_u T_k = 1 | C_k = O)\\
E_{u,O}^F &= \Pr(S_u T_k = -1 | C_k = O)\\
E_{u,\overline{O}}^T &= \Pr(S_u T_k = 1 | C_k = \overline{O})\\
E_{u,\overline{O}}^F &= \Pr(S_u T_k = -1 | C_k = \overline{O})
\end{aligned} \tag{2}$$

In addition, if source S_i is independent, I_u and J_u are defined as the probability that source S_u reports a claim C_k to be true given that claim C_k is indeed true or false. If source S_u is dependent, $I_{u,v}$ and $J_{u,v}$ are defined as the probability that source S_u reports claim C_k to be true given that source S_v also reports the claim to be true and this claim is indeed true or false. Formally, I_u, J_u, $I_{u,v}$ and $J_{u,v}$ are defined as:

$$\begin{aligned}
I_u &= \Pr(S_u C_k = 1 | C_k = T)\\
J_u &= \Pr(S_u C_k = 1 | C_k = F)\\
I_{u,v} &= \Pr(S_u C_k = 1 | S_v C_k = 1, C_k = T)\\
J_{u,v} &= \Pr(S_u C_k = 1 | S_v C_k = 1, C_k = F)
\end{aligned} \tag{3}$$

Notice that sources may report different number of claims, we denote the probability that source S_u reports a claim to be topic relevant as $tp_{u,O}$ (i.e., $tp_{u,O} = \Pr(S_u T_k = 1)$), and denote the probability that source S_u reports a claim to be topic irrelevant as $tp_{u,\overline{O}}$ (i.e., $tp_{u,\overline{O}} = \Pr(S_u T_k = -1)$). Additionally, we denote the probability that source S_u reports a claim to be true by sp_u (i.e., $sp_u = \Pr(S_u C_k = 1)$). We further denote h_O as the prior probability that a randomly chosen claim is indeed relevant to the topic of interests (i.e., $h_O = \Pr(C_k = O)$). We denote d as the prior probability that a randomly chosen claim is true (i.e., $d = \Pr(C_k = T)$). Based on the Bayes' theorem, we can obtain the relationship between the items defined above as follows:

$$\begin{aligned}
Ta_u &= \frac{E_{u,O}^T \times h_O}{tp_{u,O}}\\
Re_u &= \frac{I_u \times d}{sp_u}\\
Re_{u,v} &= \frac{I_{u,v} \times Re_v \times sp_v}{sp_u}
\end{aligned} \tag{4}$$

Finally, we define two more vectors of hidden variables Υ and Z where Υ indicates the topic relevance of claims and Z indicates the truthfulness of claims. Specifically, we define an indicator variable r_k for each claim where $r_k = 1$ when claim C_k is topic relevant and $r_k = 0$ when claim C_k is topic irrelevant. Similarly, we define another indicator variable z_k for each claim C_k where $z_k = 1$ when C_k is true and $z_k = 0$ when C_k is false.

Using the above definitions, we formally formulate the topic-aware truth discovery problem with arbitrary source dependency graphs as a multi-dimensional maximum likelihood estimation (MLE) problem: given the Source-Topic Matrix ST, the Source-Claim Matrix SC and the Social-Dependency Matrix SD, the objective is to estimate: (i) the topic relevance and truthfulness of each claim; (ii) the topic awareness and the reliability of each source. Formally, we compute:

$$\begin{aligned}
&\forall k, 1 \le k \le N : \Pr(C_k = O | ST, SC, SD)\\
&\forall k, 1 \le k \le N : \Pr(C_k = T | ST, SC, SD)\\
&\forall u, 1 \le u \le M : \Pr(C_k = O | S_u T_k = 1)\\
&\forall u, 1 \le u \le M : \Pr(C_k = T | S_u C_k = 1)
\end{aligned} \tag{5}$$

IV. Distinction from Previous Models

Before we present our estimation algorithms to solve the problem formulated in the above section, we show that our model is distinct from the the previous models in solving the truth discovery problem in social sensing [17], [28], [30], [33], [37].

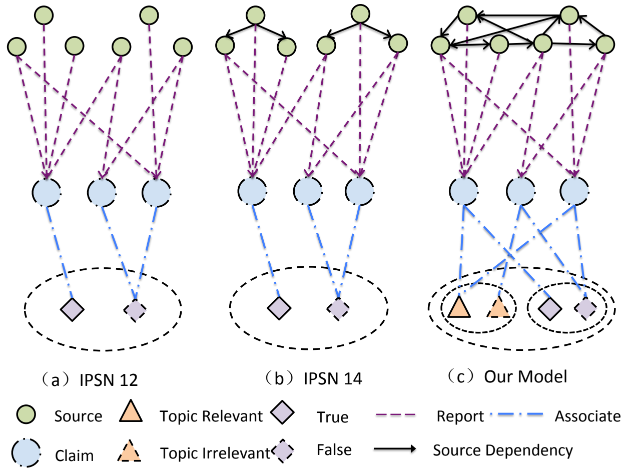


Figure 2. Model Comparison with Previous Work

Among current truth discovery models, two representative models (IPSN 12 and IPSN 14 model in Figure 2) go closest to our work. First, Wang et al. presented a basic truth discovery model (i.e., IPSN 12) that assumes all sources are independent and all claims are relevant to the topic of interests. In [30], an extended model (i.e., IPSN 14) was proposed to consider source dependencies by representing the dependency graphs as a set of disjoint trees. The extended model also assumes all claims are topic relevant. More follow-up works have been proposed to further extend the above models by considering claim dependencies [28], computation efficiency [37] and quantitative values of claims [17]. However, none of these models considered the topic relevance features of claims and arbitrary source dependency graphs, both of which are important issues to address in real world social sensing applications.

In sharp contrast to previous models, the model presented in this work explicitly incorporates the topic relevance feature of claims into the truth discovery problem and considers more general source dependencies that are represented by arbitrary graphs. In particular, as we can observe in Figure 2, we introduced two separate set of variables to represent both topic relevance and truthfulness features of claims. In our source dependency graph, any pair of sources could have a dependency link (e.g., a source could have multiple parents or cyclic dependencies in the graph). Therefore, we extend the scope of the truth discovery problem in social sensing and our estimation algorithms can solve more general problems that current solutions cannot solve. We present our estimation algorithms in the following sections.

V. Topic Relevance Identification

In this section, we present the topic relevance identification scheme: Topic-Awareness Expectation Maximization (TA-EM). The TA-EM scheme jointly estimates the topic relevance of each claim and the topic awareness of each source.

A. Deriving the Likelihood Function

EM is an optimization scheme that is commonly used to solve the MLE problem where unobserved latent variables

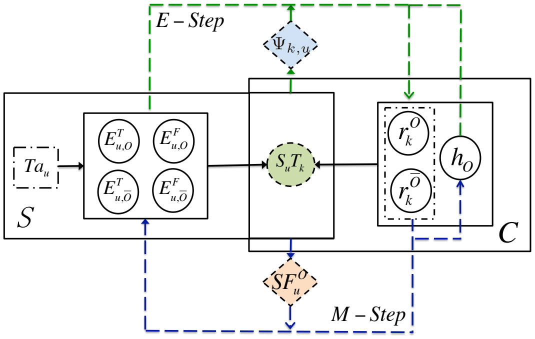


Figure 3. The E and M Steps of TA-EM Model

exist in the model [5] Specifically, it iterates between two key steps: expectation step (E-Step) and maximization step (M-step). In E-step, it computes the expectation of the log likelihood function based on the current estimates of the model parameters. In M-step, it computes the new estimates of the model parameters that maximize the expected log-likelihood function in E-step.

Given the terms and variables we defined earlier, the likelihood function $L = (\Theta_{ta}; X, \Upsilon)$ for TA-EM is as follows:

$$\begin{aligned} & L(\Theta_{ta}; X, \Upsilon) = \Pr(X, \Upsilon | \Theta_{ta}) \\ & = \prod_{k \in C} \Pr(r_k | X_k, \Theta_{ta}^{(n)}) \times \prod_{u \in S} \Psi_{k,u} \times \Pr(r_k) \end{aligned} \tag{6}$$

where $\Theta_{ta} = (E^T_{1,O}, ..., E^T_{M,O}; E^F_{1,O}, ..., E^F_{M,O}; E^T_{1,\overline{O}}, ..., E^T_{M,\overline{O}}; E^F_{1,\overline{O}}, ..., E^F_{M,\overline{O}}; h_O)$ is the vector of estimation parameters for the TA-EM scheme. X is the observed data (i.e., ST Matrix) and Υ is the latent variables. Note that $E^T_{u,O}$, $E^F_{u,O}$, $E^T_{u,\overline{O}}$, $E^F_{u,\overline{O}}$, h_O are defined in Section III. Additionally, $\Psi_{k,u}$ and $\Pr(r_k)$ are defined in Table II. In the table, $S_uT^O_k = 1$ and $S_uT^{\overline{O}}_k = 0$ when source S_u indicates claim C_k to be topic relevant. $S_uT^O_k = 0$ and $S_uT^{\overline{O}}_k = 1$ when source S_u reports claim C_k but does not indicate it to be topic relevant. $S_uT^O_k = 0$ and $S_uT^{\overline{O}}_k = 0$ when source S_u does not report claim C_k at all. Other notations are defined in the previous section. The model structure is illustrated in Figure 3.

B. The TA-EM Scheme

Given the above likelihood function, we can derive E and M steps of the proposed TA-EM scheme. First, the E-step is derived as follows:

$$\begin{aligned} & Q(\Theta_{ta} | \Theta_{ta}^{(n)}) = E_{\Upsilon | X, \Theta_{ta}^{(n)}}[logL(\Theta_{ta}; X, \Upsilon)] \\ & = \sum_{k \in C} \Upsilon(n, k) \times \sum_{u \in S} (log\Psi_{k,u} + log\Pr(r_k)) \end{aligned} \tag{7}$$

where $\Upsilon(n, k)$ is defined in Table II and n is the iteration index.

In the above table, $\Upsilon^O(n, k) = \Pr(r_k = O | X_k, \Theta_{ta}^{(n)})$. It represents the conditional probability of the claim C_j to be

Table II
NOTATIONS FOR TA-EM

$\Psi_{k,u}$	$\Pr(r_k)$	$\Upsilon(n,k)$	Constrains
$E^T_{u,O}$	h_O	$\Upsilon^O(n,k)$	$S_uT^O_k=1, S_uT^{\overline{O}}_k=0, r_k=1$
$E^F_{u,O}$	h_O	$\Upsilon^O(n,k)$	$S_uT^O_k=0, S_uT^{\overline{O}}_k=1, r_k=1$
$E^T_{u,\overline{O}}$	$1-h_O$	$1-\Upsilon^O(n,k)$	$S_uT^O_k=1, S_uT^{\overline{O}}_k=0, r_k=0$
$E^F_{u,\overline{O}}$	$1-h_O$	$1-\Upsilon^O(n,k)$	$S_uT^O_k=0, S_uT^{\overline{O}}_k=1, r_k=0$
$1-E^T_{u,O}-E^F_{u,O}$	h_O	$\Upsilon^O(n,k)$	$S_uT^O_k=0, S_uT^{\overline{O}}_k=0, r_k=1$
$1-E^T_{u,\overline{O}}-E^F_{u,\overline{O}}$	$1-h_O$	$1-\Upsilon^O(n,k)$	$S_uT^O_k=0, S_uT^{\overline{O}}_k=0, r_k=0$

topic relevant given the observed data X_k and current estimate of Θ_{ta}. $\Upsilon^O(n,k)$ can be further expressed as:

$$\begin{aligned}\Upsilon^O(n,k) &= \frac{\Pr(r_k=O;X_k,\Theta_{ta}^{(n)})}{\Pr(X_k,\Theta_{ta}^{(n)})}\\ &= \frac{L^O(n,k)\times h_O}{L^O(n,k)\times h_O + L^{\overline{O}}(n,k)\times(1-h_O)}\end{aligned} \tag{8}$$

where $L^O(n,k)$, $L^{\overline{O}}(n,k)$ are defined as:

$$\begin{aligned}L^O(n,k) &= \Pr(X_k,\Theta_{ta}^{(n)}|r_k=1)\\ &= \prod_{u=1}^{M}(E^T_{u,O})^{S_uT^O_k}\times(E^F_{u,O})^{S_uT^{\overline{O}}_k}\\ &\quad\times(1-E^T_{u,O}-E^F_{u,O})^{1-S_uT^O_k-S_uT^{\overline{O}}_k}\\ L^{\overline{O}}(n,k) &= \Pr(X_k,\Theta_{ta}^{(n)}|r_k=0)\\ &= \prod_{u=1}^{M}(E^T_{u,\overline{O}})^{S_uT^O_k}\times(E^F_{u,\overline{O}})^{S_uT^{\overline{O}}_k}\\ &\quad\times(1-E^T_{u,\overline{O}}-E^F_{u,\overline{O}})^{1-S_uT^O_k-S_uT^{\overline{O}}_k}\end{aligned} \tag{9}$$

In the M-step, we set derivatives $\frac{\partial Q}{\partial E^T_{u,O}}=0$, $\frac{\partial Q}{\partial E^F_{u,O}}=0$, $\frac{\partial Q}{\partial E^T_{u,\overline{O}}}=0$, $\frac{\partial Q}{\partial E^F_{u,\overline{O}}}=0$, $\frac{\partial Q}{\partial h_O}=0$. Solving these equations, we get expressions of the optimal $E^T_{u,O}$, $E^F_{u,O}$, $E^T_{u,\overline{O}}$, $E^F_{u,\overline{O}}$, h_O as shown in Table III. In the table, N is the total number of claims in the Source-Topic Matrix. SF^O_u is the set of claims the source S_u indicates to be topic relevant. $SF^{\overline{O}}_u$ is the set of claims the source S_u reports but does not indicate to be topic relevant.

Table III
OPTIMAL SOLUTIONS OF TA-EM

Notation	Solution	Notation	Solution
$(E^T_{u,O})^*$	$\frac{\sum_{k\in SF^O_u}\Upsilon^O(n,k)}{\sum_{k=1}^N\Upsilon^O(n,k)}$	$(E^F_{u,O})^*$	$\frac{\sum_{k\in SF^{\overline{O}}_u}\Upsilon^O(n,k)}{\sum_{k=1}^N\Upsilon^O(n,k)}$
$(E^T_{u,\overline{O}})^*$	$\frac{\sum_{k\in SF^O_u}\Upsilon^{\overline{O}}(n,k)}{\sum_{k=1}^N\Upsilon^{\overline{O}}(n,k)}$	$(E^F_{u,\overline{O}})^*$	$\frac{\sum_{k\in SF^{\overline{O}}_u}\Upsilon^{\overline{O}}(n,k)}{\sum_{k=1}^N\Upsilon^{\overline{O}}(n,k)}$
h^*_O	$\frac{\sum_{k=1}^N\Upsilon^O(n,k)}{N}$		

In summary, the input to the TA-EM scheme is the Source-Topic Matrix ST. The output is the maximum likelihood estimation of the topic relevance of claims and the topic awareness of sources. Since we assume the topic relevance feature of a claim is binary, we can classify claims as either topic relevant or topic irrelevant based on the converged value of $\Upsilon^O(n,k)$. The convergence analysis of TA-EM is presented in Section VII. Algorithm 1 shows the pseudocode of TA-EM.

Algorithm 1 Topic-Aware EM Scheme (TA-EM)

```
Initialize $\Theta_{ta}$ ($E^T_{u,O}=tp_{u,O}$, $E^F_{u,O}=0.5\times tp_{u,O}$, $E^T_{u,\overline{O}}=0.5\times$
  $tp_{u,\overline{O}}$, $E^F_{u,\overline{O}}=tp_{u,\overline{O}}$, $h_O\in(0,1)$)
$n\leftarrow 0$
repeat
    for Each $k\in C$ do
        compute $\Pr(r_k=O|X_k,\Theta_{ta}^{(n)})$ based on Equation (8)
    end for
    for Each $u\in S$ do
        compute $\Theta_{ta}^{(n)}$ based on optimal solutions which are presented in
        Table III.
    end for
    $n=n+1$
until $\Theta_{ta}^{(n)}$ converges
Let $(\Upsilon^O_k)^c$ = converged value of $\Upsilon^O(n,k)$
for Each $k\in C$ do
    if $(\Upsilon^O_k)^c\geq 0.5$ then
        consider $C_k$ as topic relevant
    else
        consider $C_k$ as topic irrelevant
    end if
end for
for Each $u\in S$ do
    calculate $Ta^*_u$ from converge values of $\Theta_{ta}$ based on Equation (4)
end for
Return the MLE on the topic relevance of claims judgment on claim $C_k$
  and the topic-awareness $Ta^*_u$ of $S_u$.
```

VI. TOPIC-AWARE TRUTH DISCOVERY WITH ARBITRARY SOURCE DEPENDENCY GRAPHS

In this section, we incorporate the TA-EM scheme from the previous section into the truth discovery problem with arbitrary source dependency graphs. We present a new scheme called Topic-Aware Source-Dependent Expectation Maximization (TASD-EM). The TASD-EM scheme jointly estimates: (i) the topic relevance and truthfulness of each claim, and (ii) the topic awareness and the reliability of each source.

A. Deriving the Likelihood Function

Given the terms and variables defined before, the likelihood function $L(\Theta_{tasd};X,\Upsilon,Z)$ for the TASD-EM scheme can be written as follows:

$$\begin{aligned}&L(\Theta_{tasd};X,\Upsilon,Z)=\Pr(X,\Upsilon,Z|\Theta_{tasd})\\ &=\prod_{k\in C}\Pr(r_k|X_k,\Theta_{tasd})\times\prod_{u\in S}\Psi_{k,u}\times\Pr(r_k)\\ &\times\Pr(z_k|X_k,\Theta_{tasd})\times\prod_{g\in G}\prod_{u\in g}\Omega_{k,g,u}\times\Pr(z_k)\end{aligned} \tag{10}$$

where $\Theta_{tasd}=(\Theta_{ta};I_1,...,I_M;J_1,...,J_M;I_{1,v},...,I_{M,v};$ $J_{1,v},...,J_{M,v};d)$ is the vector of estimation parameters for TASD-EM. Note that Θ_{ta} is defined in Section V and I_u, $I_{u,v}$, J_u, $J_{u,v}$ and d are defined in Section III. Additionally, $\Psi_{k,u}$ and $\Omega_{k,g,u}$ are defined in Table II and Table IV respectively.

In Table IV, $S_uC_k=1$ when source S_u reports claim C_k to be true and 0 otherwise. $SD_{u,v}=1$ when source S_u and

Table IV
NOTATION FOR TASD-EM

$\Omega_{k,g,u}$	Constrains
I_u	$\lvert g\rvert = 1, S_uC_k = 1, z_k = 1$
$1 - I_u$	$\lvert g\rvert = 1, S_uC_k = 0, z_k = 1$
$\prod_{v \in g} I_{u,v}$	$\lvert g\rvert > 1, S_uC_k = 1, S_vC_k = 1, SD_{u,v} = 1, z_k = 1$
$\prod_{v \in g}(1 - I_{u,v})$	$\lvert g\rvert > 1, S_uC_k = 0, S_vC_k = 1, SD_{u,v} = 1, z_k = 1$
J_u	$\lvert g\rvert = 1, S_uC_k = 1, z_k = 0$
$1 - J_u$	$\lvert g\rvert = 1, S_uC_k = 0, z_k = 0$
$\prod_{v \in g} J_{u,v}$	$\lvert g\rvert > 1, S_uC_k = 1, S_vC_k = 1, SD_{u,v} = 1, z_k = 0$
$\prod_{j \in g}(1 - J_{u,v})$	$\lvert g\rvert > 1, S_uC_k = 0, S_vC_k = 1, SD_{u,v} = 1, z_k = 0$

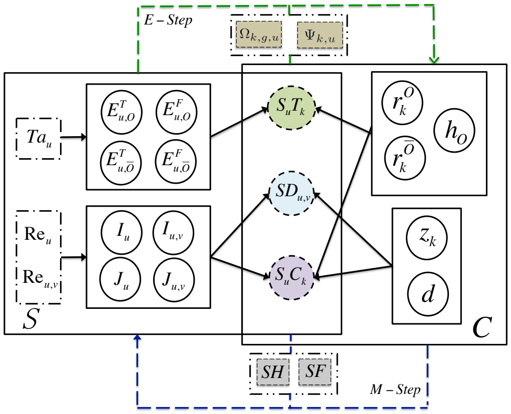


Figure 4. The E and M Steps of TASD-EM Model

source S_v have a directed dependency connection (e.g., S_v retweets S_u or replies to S_u in Twitter), and $SD_{u,v} = 0$ if S_u and S_v are independent of each other. The likelihood function represents the likelihood of the observed data (i.e., ST, SC and SD) and the values of hidden variables (i.e., Υ and Z) given the estimation parameters (i.e., Θ_{tasd}). $|g|$ denotes the size of a source dependency group g. Other notations are defined in Section III. The model structure is illustrated in Figure 4.

B. The TASD-EM Scheme

Given the above likelihood function, we can derive E and M steps of the proposed TASD-EM scheme. First, the E-step is given as follows:

$$\begin{aligned} Q(\Theta_{tasd}|\Theta_{tasd}^{(n)}) &= E_{\Upsilon,Z|X,\Theta_{tasd}^{(n)}}[logL(\Theta_{tasd};X,\Upsilon,Z)] \\ &= \sum_{k \in C}\Big\{\Upsilon(n,k) \times \sum_{u \in S}(log\Psi_{k,u} + log\text{Pr}(r_k)) \\ &+ Z(n,k) \times \sum_{g \in G}\sum_{u \in g}(log\Omega_{k,g,u} + log\text{Pr}(z_k))\Big\} \end{aligned} \tag{11}$$

$Z(n,k) = \text{Pr}(z_k = 1|X_k, \Theta_{tasd})$. It represents the conditional probability of the claim C_k to be true given the observed data X_k and current estimate of Θ_{tasd}. X_k represents the k^{th} column of the Source-Topic Matrix ST and the Source-Claim Matrix SC. $Z(n,k)$ can be further expressed as:

$$\begin{aligned} Z(n,k) &= \frac{\text{Pr}(z_k = 1; X_k, \Theta_{tasd}^{(n)})}{\text{Pr}(X_k, \Theta_{tasd}^{(n)})} \\ &= \frac{H(n,k) \times d^{(n)}}{F(n,k) \times d^{(n)} + H(n,k) \times (1 - d^{(n)})} \end{aligned} \tag{12}$$

where $F(n,k)$ and $H(n,k)$ are defined as:

$$\begin{aligned} &F(n,k) = \text{Pr}(X_k, \Theta_{tasd}^{(t)}|z_k = 1) \\ &\prod_{g \in C}\prod_{i \in g}(I_u^{S_uC_k}(1 - I_u)^{(1-S_uC_k)})^{(|g|=1)} \\ &\prod_{v \in g}((I_{u,v}^{S_uC_k\ \&\&\ S_vC_k}(1 - I_{u,v})^{(1-S_uC_k)\ \&\&\ S_vC_k)})^{SD_{u,v}})^{(|g|>1)} \\ &H(n,k) = \text{Pr}(X_k, \Theta_{tasd}^{(t)}|z_k = 0) \\ &\prod_{g \in C}\prod_{u \in g}(J_u^{S_uC_k}(1 - J_u)^{(1-S_uC_k)})^{(|g|=1)} \\ &\prod_{v \in g}((J_{u,v}^{S_uC_k\ \&\&\ S_vC_k}(1 - J_{u,v})^{(1-S_uC_k)\ \&\&\ S_vC_k)})^{SD_{u,v}})^{(|g|>1)} \end{aligned} \tag{13}$$

In the M-step, as before, we choose Θ_{tasd}^* that maximizes the $Q(\Theta_{tasd}|\Theta_{tasd}^{(n)})$ function in each iteration to be the $\Theta_{tasd}^{(n+1)}$ for the next iteration. The optimal solutions are presented in Table V. In the table, SH_u is defined as the set of claims source S_u reports to be true and $SH_{u,v}$ is the set of claims both source S_u and S_v report to be true.

Table V
OPTIMAL SOLUTIONS OF TASD-EM

Notation	Solution	Notation	Solution
$(E_{u,O}^T)^*$	$\frac{\sum_{k \in SF_u^O}\Upsilon^O(n,k)}{\sum_{k=1}^N \Upsilon^O(n,k)}$	I_u^*	$\frac{\sum_{k \in SH_u} Z(n,k)}{\sum_{k=1}^N Z(n,k)}$
$(E_{u,O}^F)^*$	$\frac{\sum_{k \in SF_u^{\overline{O}}}\Upsilon^O(n,k)}{\sum_{k=1}^N \Upsilon^O(n,k)}$	J_u^*	$\frac{\sum_{k \in SH_u}(1-Z(n,k))}{\sum_{k=1}^N (1-Z(n,k))}$
$(E_{u,\overline{O}}^T)^*$	$\frac{\sum_{k \in SF_u^O}\Upsilon^{\overline{O}}(n,k)}{\sum_{k=1}^N \Upsilon^{\overline{O}}(n,k)}$	$I_{u,v}^*$	$\frac{\sum_{k \in SH_{u,v}} Z(n,k)}{\sum_{k \in SH_v} Z(n,k)}$
$(E_{u,\overline{O}}^F)^*$	$\frac{\sum_{k \in SF_u^{\overline{O}}}\Upsilon^{\overline{O}}(n,k)}{\sum_{k=1}^N \Upsilon^{\overline{O}}(n,k)}$	$J_{u,v}^*$	$\frac{\sum_{k \in SH_{u,v}}(1-Z(n,k))}{\sum_{k \in SH_v}(1-Z(n,k))}$
h_O^*	$\frac{\sum_{k=1}^N \Upsilon^O(n,k)}{N}$	d^*	$\frac{\sum_{k=1}^N Z(n,k)}{N}$

In summary, the input to the TASD-EM scheme is the Source-Topic Matrix ST, the Source-Claim Matrix SC and the Source-Dependency Matrix SD. The output of the TASD-EM scheme is the MLE on the topic relevance and truthfulness of each claim as well as the topic-awareness and reliability of each source. The convergence analysis of TASD-EM is presented in Section VII. The pseudocode of the proposed TASD-EM scheme is shown in Algorithm 2.

VII. EVALUATION

In this section, we conduct experiments to evaluate TA-EM and TASD-EM schemes on three real-world data traces collected in the aftermath of recent emergency and disaster events. We demonstrate the effectiveness of our proposed methods on these data traces and compare the performance of our schemes to the state-of-the-art baselines. We first present the experiment settings and data pre-processing steps that were used to prepare the data for evaluation. Then we introduce the state-of-the-art baselines and evaluation metrics we used in evaluation. Finally, we show that the evaluation results demonstrate: (i) TA-EM scheme can find topic relevant claims more accurately than the compared baselines and (ii) TASD-EM can achieve non-trivial performance gains in finding more

Algorithm 2 Topic-Aware Source-Dependent EM (TASD-EM)

```
Initialize $\Theta_{tasd}$ ($E^T_{u,O} = tp_{u,O}$, $E^F_{u,O} = 0.5 \times tp_{u,O}$, $E^T_{u,\overline{O}} = 0.5 \times tp_{u,\overline{O}}$, $E^F_{u,\overline{O}} = tp_{u,\overline{O}}$, $I_u = 0.5 \times sp_u$, $I_{u,v} = 0.5$, $J_u = 0.5 \times sp_u$, $J_{u,v} = 0.5$, $h_O \in (0,1)$, $d \in (0,1)$)
$n \leftarrow 0$
repeat
    for Each $k \in C$ do
        compute $\Pr(r_k = 1|X_k, \Theta^{(n)}_{tasd})$ based on Equation (8)
        compute $\Pr(z_k = 1|X_k, \Theta^{(n)}_{tasd})$ based on Equation (12)
    end for
    for Each $u \in S$ do
        compute $\Theta^{(n)}_{tasd}$ based on optimal solutions which are presented in Table V.
    end for
    $n = n + 1$
until $\Theta^{(n)}_{tasd}$ converges
Let $(\Upsilon^O_k)^c$ = converged value of $\Upsilon^O(n,k)$
Let $(Z_k)^c$ = converged value of $Z(n,k)$
for Each $k \in C$ do
    if $(Z_k)^c \geq 0.5$ then
        consider $C_k$ as True
    else
        consider $C_k$ as False
    end if
    if $(\Upsilon^O_k)^c \geq 0.5$ then
        consider $C_k$ as topic relevant
    else
        consider $C_k$ as topic irrelevant
    end if
end for
for Each $u \in S$ do
    calculate $Ta^*_u$, $Re^*_u$, $Re^*_{u,v}$ from converge values of $\Theta_{tasd}$ based on Equation (4)
end for
Return the MLE on the topic relevance and truthfulness of claim $C_k$ as well as the topic-awareness and reliability of $S_u$.
```

valuable (i.e., relevant and truthful) claims compared to current truth discovery techniques.

A. Experimental Setups and Evaluation Metrics

1) Data Traces Statistics: In this paper, we evaluate our proposed scheme on three real-world data traces collected from Twitter in the aftermath of recent emergency and disaster events. Twitter has emerged as a new social sensing experiment platform where massive observations are uploaded voluntarily from human sensors to document the events happened in the physical world [30]. The reported observations on Twitter may be false or irrelevant to the topic of interests due to the open data collection environment and unvetted data sources [2]. However, this noisy nature of Twitter actually provides us a good opportunity to investigate the performance of the TA-EM and TASD-EM schemes in real world social sensing scenarios. In the evaluation, we selected three data traces: (i) Paris Charlie Hebdo shooting event that happened on January 7, 2015; (ii) Hurricane Arthur that happened on July 3, 2014 and (iii) Boston Marathon bombings event that happened on April 15 2013. These data traces were collected through Twitter open API using query terms and specified geographic regions related to the events. The statistics of the three data traces are summarized in Table VI.

Table VI
DATA TRACES STATISTICS

Data Trace	Paris Shooting	Hurricane Arthur	Boston Bombing
Start Date	Jan. 1 2015	July 3 2014	April 15 2013
Time Duration	3 days	3 days	4 days
Location	Paris	North Carolina	Boston
# of Tweets	39,769	27,284	63,052
# of Users Tweeted	32,391	23,106	52,583

2) Data Pre-Processing: To evaluate our methods in real-world settings, we conducted the following data pre-processing steps: (i) cluster similar tweets into the same cluster to generate claims; (ii) generate the Source-Topic Matrix (*ST* Matrix) and Source-Claim Matrix (*SC* Matrix); (iii) generate the Source Dependency Matrix (*SD* Matrix) to represent arbitrary source dependency graphs. After the above pre-processing steps, we obtained all the inputs that are needed for the proposed schemes: *ST* Matrix, *SC* Matrix and *SD* Matrix. The pre-processing steps are summarized as follows:

Clustering: we cluster similar tweets into the same cluster using a clustering algorithm based on K-means and a commonly used distance metric for micro-blog data clustering (i.e., Jaccard distance) [22]. In particular, the Jaccard distance is defined as $1 - \frac{A \cap B}{A \cup B}$, where A and B represents the set of words that appear in a tweet. Hence, the more common words two tweets share, the shorter Jaccard distance they have. We then take each Twitter user as a source and each cluster as a claim in our social sensing model described in Section III.

Source-Topic Matrix and Source-Claim Matrix Generation: we first generate the ST Matrix using the topic indicator (i.e., hashtag: #) from the tweets. In particular, if source S_u reports the claim C_k using a hashtag in the tweet, the corresponding element S_uT_k in *ST* matrix is set to 1. Similarly, if source S_u reports claim C_k without using a hashtag, the corresponding element S_uT_k is set to -1. The element S_uT_k is set to 0 when source S_u did not report claim C_k. Second, we generate the SC Matrix by associating each source with the claims he/she reported. In particular, we set the element S_uC_k in SC matrix to 1 if source S_u generates a tweet that belongs to claim (cluster) C_k and 0 otherwise.

Source Dependency Matrix Generation: we generate the Source Dependency Matrix SD based on sources' retweeting and replying behaviors on Twitter. In particular, we generated the source dependency graph as an arbitrary directed graph $G = (V, E)$ where V represents sources and E represents their dependency links. We used two heuristics to generate the links in the graph: (i) a directed edge from source S_u to source S_v is added when source S_v retweets S_u's tweets; (ii) A directed edge from source S_u to S_v is added when source S_v replies to S_u's tweets. We then constructed the Social Dependency Matrix SD by setting the corresponding element $SD_{u,v}$ to 1 when S_u has a directed link to S_v in G. We note that the above heuristics are only first approximations to estimate source dependencies from real world data. In the future, we will explore more comprehensive techniques to further refine our estimation of source dependency graphs.

3) Evaluation Metric: In our evaluation, we use the following metrics to evaluate the estimation performance of the TA-EM and TASD-EM scheme: *Precision*, *Recall*, *F1-measure* and *Accuracy*. Their definitions are given in Table VII.

Table VII
METRIC DEFINITIONS

Metric	Definition
$Precison$	$\frac{TP}{TP+FP}$
$Recall$	$\frac{TP}{TP+FN}$
$F1-measure$	$\frac{2\times Precison\times Recall}{Precison+Recall}$
$Accuracy$	$\frac{TP+TN}{TP+TN+FP+FN}$

In Table VII, TP, TN, FP and FN represents True Positives, True Negatives, False Positives and False Negatives respectively. We will further explain their meanings in the context of experiments carried out in the following subsections.

B. Evaluation of Our Methods

In this subsection, we evaluate the performance of the proposed TA-EM and TASD-EM scheme and compare them to the state-of-the-art truth discovery methods.

1) Evaluation on Topic Relevance Identification: We first evaluate the capability of TA-EM scheme to correctly identify the topic relevant claims from noisy social sensing data. We compared the TA-EM with several baselines. The first one is *Voting*: it simply assumes the topic relevance of a claim is reflected by the number of times it is repeated on Twitter: the more repetitions of a claim, the more likely it is relevant to a topic of interests. The second baseline is the *Hashtag*: it considers a claim to be topic relevant if the claim contains the hashtag related to the specified topic. The third baseline is the *HIST* (Hyperlink-Induced Topic Search) [15]: it assumes a linear relationship between the source's topic awareness and the claim's topic relevance. The last baseline is the *TruthFinder* [39]: it can estimate the topic relevance of a claim using a heuristic based pesudo-probabilistic model.

In our evaluation, the outputs of the above schemes were manually graded to determine their performance on topic relevant claim identification. Due to man-power limitations, we generated the evaluation set by taking the union of the top 50 relevant claims returned by each scheme to avoid possible sampling bias towards any particular scheme. We collected the ground truth of the evaluation set using the following rubric:

- Topic Relevant Claims: claims that describe a physical or social event which is clearly related with a chosen topic (e.g., Paris Shooting, Hurricane Arthur, or Boston Bombing in our selected datasets).
- Topic Irrelevant Claims: claims that do not meet the definition of the topic relevant claims.

In our evaluation, the True Positives and True Negatives are the claims that are correctly classified by a particular scheme as topic relevant and irrelevant ones respectively. The False Positives and False Negatives are the irrelevant and relevant claims that are misclassified to each other respectively.

The evaluation results of Paris Shooting data trace are shown in Figure 5. We can observe that *TA-EM* outperforms the compared baselines in all evaluation metrics. The largest performance gain achieved by *TA-EM* on F1-measure and accuracy over the best performed baseline (i.e., *Hashtag*) are 10% and 20% respectively. The results of Hurricane Arthur data trace are presented in Figure 6. *TA-EM* continues to outperform all baselines and the largest performance gain achieved by *TA-EM* on F1-measure and accuracy is 13% and 16% respectively. The results of Boston Bombing data trace are similar and are not presented due to space limit.

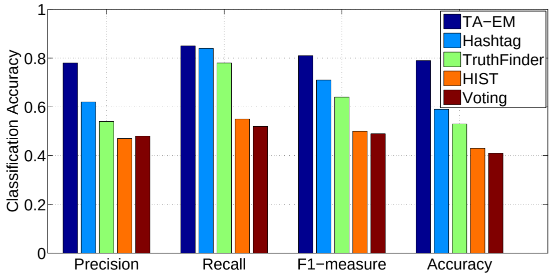


Figure 5. Topic Relevance Identification on Paris Shooting Trace

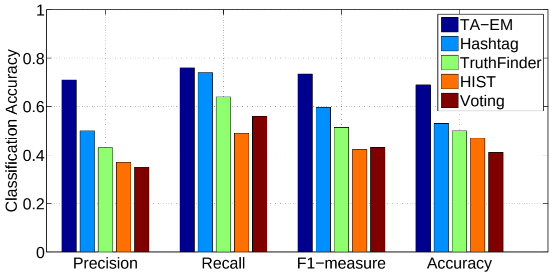
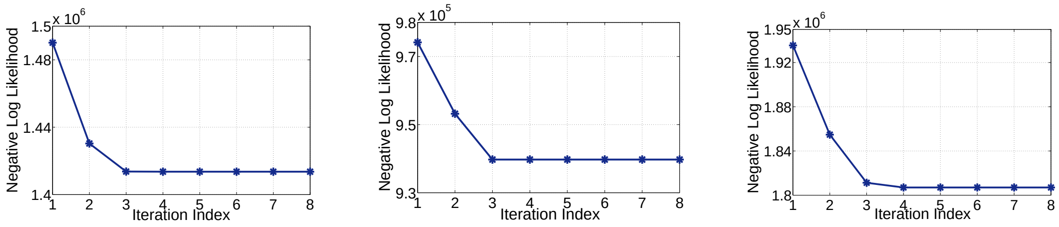


Figure 6. Topic Relevance Identification on Hurricane Arthur Trace

We also perform the convergence analysis of the TA-EM scheme and the results are presented in Figure 7. We observe the TA-EM scheme converges within a few iterations on all three data traces. The encouraging results from the real world data traces demonstrate the effectiveness of using TA-EM scheme to correctly identify the topic relevant claims from noisy social sensing data.

2) Estimation Performance on Topic-Aware Truth Discovery: In the evaluation of *TASD-EM* scheme, we consider four variants of the *TASD-EM*: (i) *SD-EM-RT*: a simplified version of TASD-EM scheme which does not include the topic relevance identification of claims. The source dependency graphs are constructed by using the retweet relationship between sources; (ii) *SD-EM-Re*: it is similar as the SD-EM-RT scheme but the source dependency graphs are constructed by using the reply relationship between sources; (iii) *TASD-EM-RT*: the full version of TASD-EM scheme that considers both topic relevance identification of claims and the source dependency graphs built by using the retweet relationship; (iv) *TASD-EM-Re*: it is similar as *TASD-EM-RT* but the source dependency graphs are built by using the reply relationship.

We compare the TASD-EM scheme and its variants with the state-of-the-art truth discovery solutions in social sensing



(a) Paris Shooting Trace (b) Hurricane Arthur Trace (c) Boston Bombing Trace

Figure 7. Convergence Rate of TA-EM

literature. The first one is *IPSN12* [33]: it computes the claims' truthfulness and sources' reliability in an iterative way and has been shown to outperform four fact-finding techniques in identifying truthful claims from social sensing data. However, it assume sources are all *independent*. The second baseline is *IPSN14* [30], it extends *IPSN12* by explicitly considering the social dependency graphs as a set of disjoint trees. In our evaluation, we also consider two variants of *IPSN14*: *IPSN14-RT* (i.e., using the retweet relationship to build source dependency graphs) and *IPSN14-Re* (i.e., using the reply relationship to build source dependency graphs). We evaluate the performance of our proposed scheme and its variants (i.e., *TASD-EM-RT*, *TASD-EM-Re*, *SD-EM-RT* and *SD-EM-Re*) and compare them with the discussed truth discovery schemes and a few other state-of-art baselines (i.e., *IPSN14-RT*, *IPSN14-Re*, *Regular-EM*, *TruthFinder*, *HIST* and *Voting*).

To incorporate both topic relevance and truthfulness of claims into our evaluation, we generalized the concept of a *truthful* claim from the truth discovery problem to a *valuable* claim in the topic-aware truth discovery problem. In particular, a valuable claim is defined as a claim that is both truthful and relevant to the specified topic of interests. The valuable claims are the ones that are eventually useful in the decision making process. Similarly as the topic relevance identification evaluation, we generated the evaluation set by taking the union of the top 50 claims returned by different schemes. We collected the ground truth of the evaluation set using the following rubric:

- Valuable Claims: Claims that are statements of a physical or social event, which is related to the selected topic (i.e., Paris Shooting, Hurricane Arthur, or Boston Bombing) and generally observable by multiple independent observers and corroborated by credible sources external to Twitter (e.g., mainstream news media).
- Unconfirmed Claims: Claims that do not satisfy the requirement of valuable claims.

We notice that unconfirmed claims may include the valueless claims and some possibly valuable claims that cannot be independently verified by external sources. Hence, our evaluation provides pessimistic performance bounds on the estimation results by taking the unconfirmed claims as valueless. The True Positives and True Negatives in this experiment are the claims that are correctly classified by a particular scheme as valuable and valueless ones respectively. The False Positives and False Negatives are the valueless and valuable claims that are misclassified to each other respectively.

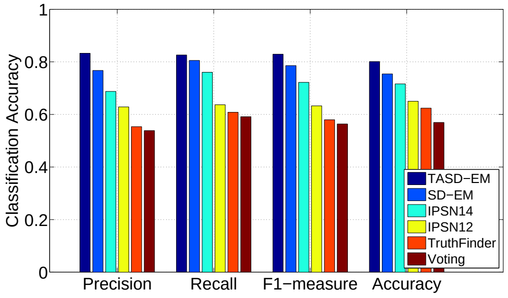


Figure 8. Estimation Results of Truth Discovery on Paris Shooting Trace

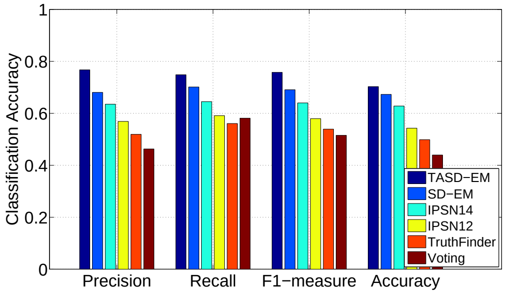
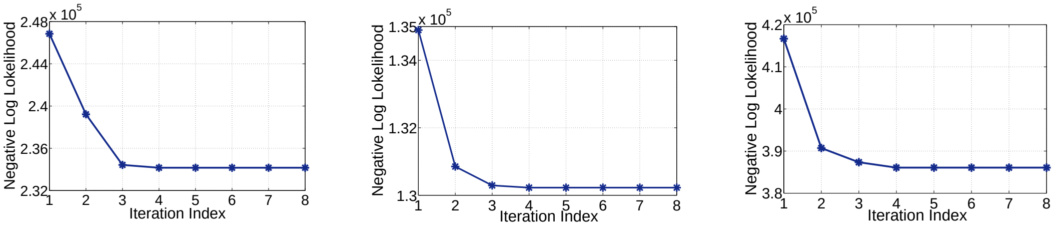


Figure 9. Estimation Results of Truth Discovery on Hurricane Arthur Trace

The evaluation results of Paris Shooting data trace are shown in Figure 8. We only showed the best performed variant of a scheme in the Figure (e.g., *TASD-EM* is actually *TASD-EM-RT*). The full evaluation results of all schemes are presented in Table VIII. We observe that the proposed schemes (i.e., *TASD-EM*, *SD-EM*) outperform all baselines. Specifically, the largest performance gain achieved by *TASD-EM* compared to the best performed baseline (i.e., *IPSN14*) on precision, recall, F1-measure and accuracy is 14%, 7%, 10% and 9% respectively. The results on Hurricane Arthur data trace are shown in Figure 9 and Table IX. We observe that our *TASD-EM* continues to outperform the compared baselines and the largest performance gain it achieved over the best performed baseline on precision, recall, F1-measure and accuracy is 13%, 10%, 11% and 7% respectively. We do present them here due to space limit. The performance improvements of *TASD-EM* are achieved by explicitly considering the *topic relevance* feature of claims and *arbitrary source dependency graphs* in social sensing, which are missing from the state-of-the-art solutions. Finally, we perform the convergence analysis of the TASD-EM scheme on the three data traces and the results are presented in Figure 10. We observe the TASD-EM scheme converges quickly on all data traces.



(a) Paris Shooting Trace (b) Hurricane Arthur Trace (c) Boston Bombing Trace

Figure 10. Convergence Rate of TASD-EM

Table VIII
EVALUATION RESULTS ON PARIS SHOOTING DATA TRACE

Method	Precision	Recall	F1-measure	Accuracy
TASD-EM-RT	**0.8327**	**0.8261**	**0.8293**	**0.8012**
TASD-EM-Reply	0.8063	0.8227	0.8144	0.7875
SD-EM-RT	0.7672	0.8054	0.7858	0.7541
SD-EM-Reply	0.7278	0.7732	0.7498	0.7367
IPSN14-RT	0.6875	0.7604	0.7221	0.7162
IPSN14-Reply	0.6594	0.7282	0.6921	0.6643
IPSN12	0.6287	0.6369	0.6327	0.6502
TruthFinding	0.5535	0.6083	0.5796	0.6241
HIST	0.5784	0.6027	0.5903	0.5816
Voting	0.5387	0.5914	0.5638	0.5695

Table IX
EVALUATION RESULTS ON HURRICANE ARTHUR TRACE

Method	Precision	Recall	F1-measure	Accuracy
TASD-EM-RT	**0.7633**	**0.7485**	**0.7558**	**0.7029**
TASD-EM-Reply	0.7381	0.7096	0.7235	0.6884
SD-EM-RT	0.6805	0.7013	0.6907	0.6728
SD-EM-Reply	0.6585	0.6728	0.6656	0.6357
IPSN14-RT	0.6352	0.6489	0.6419	0.6281
IPSN14-Reply	0.6128	0.6287	0.6206	0.5983
IPSN12	0.5687	0.5910	0.5796	0.5428
TruthFinding	0.5192	0.5607	0.5392	0.49858
HIST	0.4815	0.5795	0.5259	0.4716
Voting	0.4627	0.5812	0.5152	0.4395

VIII. DISCUSSIONS AND LIMITATIONS

In this paper, we focus on a single general topic of the selected event (e.g., Paris Shooting, Hurricane Arthur, or Boston Bombing). However, subtopics could also exist under the general topic of the event. For example, in the aftermath of an earthquake, people may report their observations on different aspects of the disaster (e.g., damage, rescue, resource allocation, disease spread). Each aspect of the event can be considered as a subtopic. Our analytical framework can be easily extended to handle multiple *independent* features of claims. For example, we can apply a simple extension of TA-EM to incorporate multiple independent subtopics of an event by assigning a vector of hidden variable for each subtopic. This task becomes more challenging when the subtopics are *dependent* (e.g., claims reporting rescue progress can be related to claims reporting the damage in the same area). Recent work in social sensing has made good progress to address the dependencies between claims [28]. We can borrow insights from the previous work on claim dependency to handle dependent subtopics of claims in an extended framework.

It would also be interesting to investigate the possibility of developing a fully Bayesian approach to solve the truth discovery problem in social sensing. This approach will enable additional indications of confidence on the truth discovery results, which could be tremendously useful for the analysts and decision makers to understand the uncertainties in the analysis results. In a different line of work, we have started to develop accuracy bounds to rigorously quantify the quality of the truth discovery results in social sensing using estimation theoretical approaches [32]. We will further explore the possibility of using a fully Bayesian approach to address the uncertainty problem.

Finally, we should note that the topic aware social sensing model with arbitrary source dependency graphs developed in this paper is not only applicable to applications based on Twitter. It can be also applied to a much broader set of social sensing applications, where the data are collected from both human sensors or the devices on their behalf. Examples include reporting the number of available fitness machines in a gym, reporting locations of potholes on city streets and reporting the invasive species in a national park. In these applications, humans sources could also generate irrelevant measurements (e.g., by misidentifying the target objects or performing the wrong operations) and forward unverified reports they received from other sources. Our solution can be used to address similar truth discovery and source dependency problems in these applications.

IX. CONCLUSION

This paper develops a multi-dimensional maximum likelihood estimation framework to solve the topic-aware truth discovery problem with arbitrary source dependency graphs in social sensing applications. The framework explicitly incorporates both the topic relevance feature of claims and arbitrary source dependencies into the truth discovery solutions. The proposed approach jointly estimates the topic awareness and reliability of sources as well as the topic relevance and truthfulness of claims using expectation maximization schemes. We evaluated our solution (i.e., TA-EM and TASD-EM scheme) using three real world data traces collected from Twitter. The results showed that our solution achieved non-trivial performance gains in correctly identifying topic relevant and truthful claims compared to the state-of-the-art baselines. The results of the paper is important because it lays out a solid analytical foundation to explore both the topic relevance feature of claims and arbitrary source dependency graphs in social sensing applications using a principled approach.

ACKNOWLEDGMENT

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

REFERENCES

[1] T. Abdelzaher and D. Wang. Analytic challenges in social sensing. In *The Art of Wireless Sensor Networks*, pages 609–638. Springer, 2014.

[2] C. C. Aggarwal and T. Abdelzaher. Social sensing. In *Managing and Mining Sensor Data*, pages 237–297. Springer, 2013.

[3] L. Cabral and A. Hortacsu. The dynamics of seller reputation: Evidence from eBay. *The Journal of Industrial Economics*, 58(1):54–78, 2010.

[4] H.-C. Chang. A new perspective on twitter hashtag use: diffusion of innovation theory. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–4, 2010.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

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

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

[8] M. Gupta and J. Han. Heterogeneous network-based trust analysis: a survey. *ACM SIGKDD Explorations Newsletter*, 13(1):54–71, 2011.

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

[10] 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.

[11] C. Huang and D. Wang. Time-aware truth discovery in social sensing. In *Mobile Ad Hoc and Sensor Systems (MASS), 2015 IEEE 12th International Conference on*, pages 479–480. IEEE, 2015.

[12] K. L. Huang, S. S. Kanhere, and W. Hu. Are you contributing trustworthy data?: the case for a reputation system in participatory sensing. In *Proceedings of the 13th ACM international conference on Modeling, analysis, and simulation of wireless and mobile systems*, pages 14–22. ACM, 2010.

[13] K. L. Huang, S. S. Kanhere, and W. Hu. On the need for a reputation system in mobile phone based sensing. *Ad Hoc Networks*, 12:130–149, 2014.

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

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

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

[17] 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.

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

[19] A. Patil, J. Liu, and J. Gao. Predicting group stability in online social networks. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1021–1030. International World Wide Web Conferences Steering Committee, 2013.

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

[21] G.-J. Qi, C. C. Aggarwal, J. Han, and T. Huang. Mining collective intelligence in diverse groups. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1041–1052. International World Wide Web Conferences Steering Committee, 2013.

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

[23] X. Sheng and Y.-H. Hu. Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks. *Signal Processing, IEEE Transactions on*, 53(1):44–53, 2005.

[24] M. Uddin, M. Amin, H. Le, T. Abdelzaher, B. Szymanski, and T. Nguyen. On diversifying source selection in social sensing. In *Networked Sensing Systems (INSS), 2012 Ninth International Conference on*, pages 1 –8, june 2012.

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

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

[27] 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.

[28] 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.

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

[30] 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.

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

[32] D. Wang, L. Kaplan, T. Abdelzaher, and C. C. Aggarwal. On scalability and robustness limitations of real and asymptotic confidence bounds in social sensing. In *The 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON 12)*, June 2012.

[33] 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.

[34] H. Wang, D. Lymberopoulos, and J. Liu. Local business ambience characterization through mobile audio sensing. In *Proceedings of the 23rd international conference on World wide web*, pages 293–304. ACM, 2014.

[35] J. Wang, Y. Zhao, and D. Wang. A novel fast anti-collision algorithm for rfid systems. In *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on*, pages 2044–2047. IEEE, 2007.

[36] J.-w. WANG, D. WANG, K. TIMO, and Y.-p. ZHAO. A novel anti-collision protocol in multiple readers rfid sensor networks [j]. *Chinese Journal of Sensors and Actuators*, 8:026, 2008.

[37] S. Wang, L. Su, S. Li, S. Hu, T. Amin, H. Wang, S. Yao, L. Kaplan, and T. Abdelzaher. Scalable social sensing of interdependent phenomena. In *Proceedings of the 14th International Conference on Information Processing in Sensor Networks*, pages 202–213. ACM, 2015.

[38] Y. Wang and J. Vassileva. A review on trust and reputation for web service selection. In *Distributed Computing Systems Workshops, 2007. ICDCSW'07. 27th International Conference on*, pages 25–25. IEEE, 2007.

[39] 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.

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

[41] D. Yu, H. Huang, T. Cassidy, H. Ji, C. Wang, S. Zhi, J. Han, C. Voss, and M. . Magdon-Ismail. The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding. In *The 25th International Conference on Computational Linguistics (COLING)*, 2014.

[42] P. Zhang and A. Purohit. The cloud meets the crowd: Framework for distributed cloud sensing. In *UbiComp*, 2011.

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