Critical Source Selection in Social Sensing Applications

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

***Abstract*—Social sensing has emerged as a new data collection paradigm in networked sensing applications where humans are used as "sensors" to report their observations about the physical world. While many previous studies in social sensing focus on the problem of ascertaining the reliability of data sources and the correctness of their reported claims (often known as *truth discovery*), this paper investigates a new problem of *critical source selection*. The goal of this problem is to identify a subset of critical sources that can help effectively reduce the computational complexity of the original truth discovery problem and improve the accuracy of the analysis results. In this paper, we propose a new scheme, *Critical Sources Selection (CSS)* scheme, to find the critical set of sources by explicitly exploring both *dependency* and *speak rate* of sources. We evaluated the performance of our scheme and compared it to the state-of-the-art baselines using two data traces collected from a real world social sensing application. The results showed that our scheme significantly outperforms the baselines by finding more truthful information at a faster speed.**

***Keywords*-Source Selection, Source Dependency, Speak Rate, Social Sensing, Twitter**

I. Introduction

This paper develops a new scheme to solve the critical source selection problem in social sensing applications. Social sensing has emerged as a new networked sensing paradigm of collecting observations about the physical environment from humans or devices on their behalf. This paradigm is motivated by the proliferation of digital sensors in the possession of common individuals (e.g., smartphones) and the wide adaptation of online social media (e.g., Twitter, Facebook). In social sensing applications, people can report certain observations of their environment, such as traffic condition at various locales [7], pothole information on streets [19] and available gas stations in the aftermath of a disaster [34]. One key challenge of using "humans as sensors" is to estimate the correctness of observations (i.e., *claims*) and the reliability of data sources without knowing either of them *a priori*. We refer to this problem as *truth discovery problem*.

In this paper, we study a new problem of *critical source selection* where the goal is to identify a subset of critical sources that can help effectively reduce the computational complexity of the original truth discovery problem and improve the accuracy of the analysis results. First, it is critical to consider the source dependency in solving this problem.

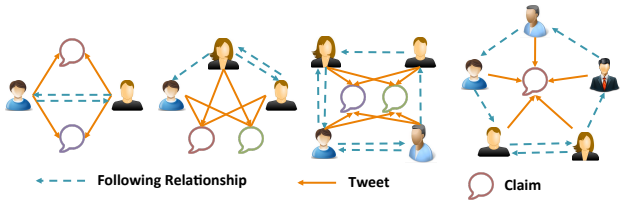

Figure 1. Source Dependency Examples on Twitter

In social sensing, it is not unusual for a human source to forward claims they received from others (e.g., friends from their social networks) [39]. Figure 1 shows some simple examples extracted from real-word Twitter data where sources with social connections (i.e., following relationship) report the same claim. From a networked sensing perspective, such dependency between sources can easily introduce correlation and redundancy between reported observations, which are shown to affect truth discovery results negatively if they are not appropriately modeled [34]. Previous works [12], [34], [25] have started to consider the source dependency between sources in truth discovery tasks by partitioning them into independent groups where sources in different groups are considered to be independent. However, the complexity of their solutions grows exponentially with respect to the maximum size of the independent group, making them impractical in many large-scale social sensing applications [12]. In this paper, we develop a new source selection scheme to explicitly consider the source dependency in the source selection process.

In addition to the source dependency, the speak rate of a source (i.e., how chatty a source is) is another important factor to consider in the critical source selection solution. In social sensing, different sources often report different number of claims (e.g., some sources are more chatty than others). The speak rate of a source has been found to has a strong positive correlation with both the accuracy and the granularity of the source reliability estimation, which also directly affects the estimation of the claim correctness [18]. Therefore, the goal of our critical sensor selection scheme is to (i) maximize the average speak rate of the selected sources and (ii) minimize the dependency between them. However, those two objectives can be at odds with each

other, which makes the critical sensor selection problem a non-trivial problem to solve.

Previous work has made significant progress to study the problem of source selection in sensor network and data fusion [32], [27], [6], [10]. However, most of current solutions either ignore the source dependency or the speak rate in their models, which have led to suboptimal source selection results by selecting redundant sources or sources with inaccurate source reliability estimations. In this paper, we present a Critical Source Selection (CSS) scheme that explicitly incorporates both the *source dependency* and the *speak rate* feature into the critical source selection process. In particular, we formulate our critical source selection problem as a constraint optimization problem with multiple objectives and develop an efficient algorithm to solve it. We evaluated our CSS scheme in comparison with the state-of-the-art baselines using two real-world social sensing data traces collected from Twitter (i.e., one for Paris Attack event and the other for Oregon Shooting event, both in 2015). The results showed that our scheme significantly outperforms the baselines by finding more truthful information at a faster speed.

In summary, **our contributions** are as follows:

- In this paper, we investigate the problem of critical source selection in social sensing to reduce the complexity of a truth discovery problem and improve the accuracy of estimation results at the same time.
- We develop a new approach (i.e., CSS scheme) that selects a critical set of sources by exploring both the source dependency and their speak rates.
- We perform extensive experiments to compare the performance of our CSS scheme with the-sate-of-the-art baselines using real-world social sensing data. The evaluation results demonstrate the effectiveness and efficiency of our scheme.

The rest of this paper is organized as follows: we discuss the related work in Section II. In Section III, we present the problem of critical source selection. The proposed critical source selection scheme is discussed in Section IV. Experiment and evaluation are presented in Section V. Finally, we conclude the paper in Section VI.

II. Related Work

Social Sensing. Social sensing has emerged as a new sensing paradigm which attracted much attention in sensor networks research [1]. The ideas of getting people involved into the loop of the sensing process (e.g., participatory [4], [35], opportunistic [17], [11] and human-centric sensing [9], [37]) have been extensively studied in projects such as MetroSense [5], Urban Sensing [16] and SurroundSense [3]. The idea of using humans as sensors themselves came more recently [30]. For example, human sensors can contribute their observations through "sensing campaigns" [26], [13] or social data scavenging [36], [40]. Current works in social sensing have addressed many important challenges such as privacy perseverance [38], incentives design [28], scalability [15], [41] and social interaction promotions [31]. However, the source selection remains to be a critical and open research question in social sensing. In this work, we study the problem of *critical source selection* to reduce the computational complexity of the original truth discovery problem and improve the accuracy of the analysis results.

Truth Discovery in Social Sensing. Data quality and trustworthiness is one of the fundamental challenges in social sensing. Prior works in social sensing have made significant advances to infer the credibility of reported data [24], [12], [23], [33], [34], [14]. For example, Ouyang et al. [24] investigated the potential of leveraging crowds as sensors to detect the true value of quantitative characteristics from noisy social sensing data. Huang et al. explored the topic relevance of claims and arbitrary source dependency problem in social sensing and developed a topic-aware truth discovery solution [12]. Meng et al. developed a truth discovery scheme that considers the correlated claims by modeling claims' correlations as regularization terms [23]. Zhao et al. studied the problem of real-time truth discovery and developed a probabilistic model to efficiently handle the streaming data [42]. Marshall et al. investigated the semantics of the claims in the truth discovery solutions of social sensing [20], [22], [21]. Wang at al. considered source dependency by assuming that it can be represented by sets of disjoint trees [34]. All the above works solve the *truth discovery problem* and focus on modeling the relationship between source reliability and claim correctness. In contrast, this paper solves a new problem of *critical source selection* which can help improve both effectiveness and efficiency of the above truth discovery solutions.

Source Selection in Social Sensing. There exists a good amount of work on the topic of *source selection* in networked sensing, data mining and data base communities [32], [27], [6], [2], [10]. For example, Uddin et al. investigated the problem of diversifying the source selection in social sensing based on the social connections between sources. Rekatsinas et al. [27] studied the problem of source selection for dynamic sources whose contents change over time. Dong et al. [6] proposed an algorithm to select a subset of sources in data fusion applications by considering integration cost. Hosseini et al. selected the subset of data sources to predict the state of all other sources by considering source correlations [10]. Amintoosi et al. [2] proposed a a privacy-aware participant selection framework that explicitly protects user's privacy in the social sensing applications. However, most of current solutions either ignore the source dependency or the speak rate in their models. In contrast, this paper explicitly incorporates both the *source dependency* and *speak rate* into the critical source selection process.

III. Problem Formulation

We consider a social sensing scenario where a set of X sources (denoted as S) who jointly report a set of Y claims (denoted as C). We denote an individual source as $S_i \in S,\ i \in [1, ..., X]$ and an individual claim as $C_j \in C,\ j \in [1, ..., Y]$, where i and j are the source and claim index respectively. The same claim can be made by multiple sources and each source can report multiple claims. We define the following terms we will use in our problem formulation.

Definition 1. ***Source-Claim Matrix*** SC. *We define Source-Claim Matrix* $\boldsymbol{SC}_{X\times Y}$ *to represent whether a source reports a claim or not. In particular, in* SC*, we set* $SC_{i,j} = 1$ *if source* S_i *reports claim* C_j *and* $SC_{i,j} = 0$ *otherwise.*

Definition 2. ***Speak-Rate Vector*** SR. *We define Speak-Rate Vector* $\boldsymbol{SR}_i$ *to represent* how chatty a source is. *Specifically, the element* SR_i *in* SR *is the number of claims reported by source* S_i *normalized by the total number of claims:* $SR_i = \frac{\Sigma_{j=1}^{Y} SC_{i,j}}{Y}$.

Definition 3. ***Source-Dependency-Score Matrix*** SDS. *We define Source-Dependency-Score Matrix* $\boldsymbol{SDS}_{X\times X}$ *to represent* dependency between each pair of sources. *Specifically, the element* $SDS_{i,i'}$ *in* SDS *is the number of common claims reported by both source* S_i *and* $S_{i'}$.

We summarize the defined notations in Table I.

Table I
Summary of Notations

Symbol	Interpretation
S	set of sources
C	set of claims
SC	source-claim matrix
SR	speak-rate vector
SDS	source-dependency-score matrix

In social sensing applications, the estimation accuracy of source reliability is positively correlated with the speak rate of data sources [36]. The first objective of our critical source selection problem is to maximize the speak rates of the set of selected sources. Furthermore, observations from *independent sources* often provide more critical information to solve the truth discovery problem [34]. In Definition 3, we use the number of commonly reported claims to measure the dependency between two sources. This is based on the assumption that two independent sources are less likely to report many claims in common [12]. Therefore, the second objective here is to minimize the dependency between the selected sources. Finally, the claims reported by the selected sources should cover all claims in C for the completeness of the problem.

With the above definitions, we can formulate the *critical source selection* problem as follows: given the Source-Claim Matrix $\boldsymbol{SC}_{X\times Y}$, Speak-Rate Vector $\boldsymbol{SR}_X$ and Source-Dependency Matrix $\boldsymbol{SDS}_{X\times X}$, the goal is to select the set of critical sources (denoted by S^*) whose reported claims can cover the claim set C by maximizing their total speak rates and minimizing their total dependency scores. Formally, the problem can be represented as follows:

$$
\begin{aligned}
& max \sum_{i=1}^{X} CV_i \cdot \delta_i \\
& min \sum_{i=1}^{X} \sum_{i' \neq i} SDS_{i,i'} \cdot \delta_i \cdot \delta_{i'} \\
& s.t.\ \ \delta_i \in 0, 1, \quad i = 1, ..., X \\
& \qquad \bigcup C_{S_i} = C, \quad S_i \in S
\end{aligned} \tag{1}
$$

where $\delta_i = 1(\text{or } 0)$ indicates that source S_i is selected (or not).

IV. Source Selection

In the previous section, we formulate the *critical source selection* problem as a constraint optimization problem. One possible solution to the optimization problem is to perform brute-force search. However, the time complexity of brute-force search is $O(2^{|S|})$ ($|S|$ is the number of sources), which is not scalable in many practical social sensing applciations. There we need to develop a more efficient solution. In the rest of this section, we first prove that the formulated critical source selection problem is NP-hard. We then present the details of our CSS scheme and summarize it using a pesudocode.

A. Complexity Analysis of the Formulated Problem

In this subsection, we prove the formulated problem is a NP-hard problem. Based on the definitions in Section III, we construct a graph $G = (S, C; E_S, E_{SC})$ based on the Source-Claim Matrix SC, Source-Dependency-Score Matrix SDS and Speak-Rate Vector SR as follows:

- A source S_i represents a vertex in S;
- A claim C_j represents a vertex in C;
- E_S is the set of edges between the vertices of S to represent the dependency between sources in SDS. In particular, if the element $SDS_{i,i'} > 0$, we have an edge between source S_i and $S_{i'}$.
- E_{SC} is the set of edges between vertices of S and C to represent report behaviors in SC. Specifically, if the element $SC_{i,j}$ in SC is 1, we have a edge between source S_i and claim C_j.
- We define wv_i as SR_i (i.e., speak rate of source S_i) to represent the vertex weight of vertex S_i and $we_{i,i'}$ as $SDS_{i,i'}$ (i.e., dependence score between source S_i and $S_{i'}$) to represent the edge weight between vertex S_i and $S_{i'}$.

- We further define two weight functions $w_{E_S} : E_S \mapsto R^+$ and $w_S : S \mapsto R^+$ to represent the dependency scores between sources and speak rates of sources respectively.

The objective is to find a subset S^* of S such that every vertex in C is connected to the vertex in S^* and satisfies the following objective functions:

i) the sum of vertex weights in S^* (i.e., $\Sigma\ wv_i;\ S_i \in S^*$) is maximized;

ii) the sum of edge weights $we_{i,i'}$ in the subgraph induced by S^* (i.e., $\Sigma\ we_i;\ S_i, S_{i'} \in S^*$) is minimized;

We first consider a simplified version of the above problem by only considering the objective of minimizing the total dependency scores of selected sources. If we can prove that this simplified version is NP-hard, the original version is also NP-hard. We formally define the decision version of the simplified problem as follows:

Definition 4. *Given a graph $G = (S, C; E_S, E_{SC})$, a weight function $w_{E_S} : E_S \mapsto R^+$, a weight function $w_S : S \mapsto R^+$, and a positive number k, where S and C are two sets of vertices. E_S is a set of edges only among the vertices of S. E_{SC} is a set of edges between vertices S and C. The objective is to decide whether there is a subset S^* of S such that every vertex in C is connected to the vertex in S^* and the sum of edge weights in the subgraph induced by S^* is at most k.*

To prove that the simplified version is NP-hard, we need to demonstrate that the decision version is NP-complete. After that, we can conclude that the problem formulated in Equation (1) is NP-hard.

B. The Critical Source Selection Scheme

The proof in above section shows that the formulated problem is NP-hard, we need to develop an efficient solution. In this work, we propose the *Critical Source Selection (CSS)* scheme to select the critical set of sources.

Based on the problem formulation in Equation (1), there are two objectives: (i) maximize the speak rates of the selected sources; (ii) minimize the dependency scores between the selected sources. We take a common approach in optimization and convert multi-objective programming to single-objective programming using linear combination [43]. We can rewrite Equation (1) as:

$$max \sum_{i=1}^{X} SR_i \cdot \eta_i - \varphi \cdot \sum_{i=1}^{X} \sum_{i' \neq i} SDS_{i,i'} \cdot \eta_i \eta_{i'}$$
$$s.t.\ \ \eta_i \in 0, 1, \quad i = 1, ..., X$$
$$\bigcup C_{S_i} = C, \quad S_i = 1 \text{ and } S_i \in S \quad (2)$$

where φ is a parameter to balance our two objectives.

We denote a graph $G_s = (S, E_S, W_e, W_v)$, where W_e, W_v represent the set of edge weights and the set of vertex weights respectively. Without loss of generality, we use v_i, $e_{i,i'}$, $we_{i,i'}$, wv_i to represent the vertex, edge, edge weight and vertex weight respectively ($i \in [1, ..., X]$, $i' \in [1, ..., X]$ and $i \neq i'$). In this work, v_i is source S_i. $e_{i,i'}$ represents the dependency relationships between source S_i and $S_{i'}$. We further define Ne_i as the vertices which are connected to vertex v_i and t as the iteration index.

In particular, we first construct a set S^* and C^* to contain the selected sources and the set of claims reported from the selected sources. The key steps of CSS scheme is summarized as:

- We initialize S^* and C^* as $\emptyset$.
- We do the following three sub-steps iteratively:
  i) We select the vertex in graph G_s with the largest vertex weight wv_i. Without loss of generality, we denote the selected node as v_i.
  ii) We conduct vertex weight updates on other vertices which connected to vertex v_i. Specifically, the weight $wv_{i'}$ on vertex $v_{i'}$ ($i' \in Ne_i$) is updated as $wv_{i'} = wv_{i'} - we_{i,i'} \cdot \varphi$. Here, we update the vertex weight of the connected vertices of v_i by balancing the two objectives in Equation (2).
  iii) We add vertex v_i to the set of selected sources S^* and remove it from graph G_s together with all the edges connected to vertex v_i.

Figure 2 shows a simple illustrative example for CSS Scheme. In the source selection process, we firstly select vertex v_5 with the largest vertex weight. After that, we update the vertex weights of the vertices connected to v_5 and remove v_5 as well as the corresponding edges from the current graph.

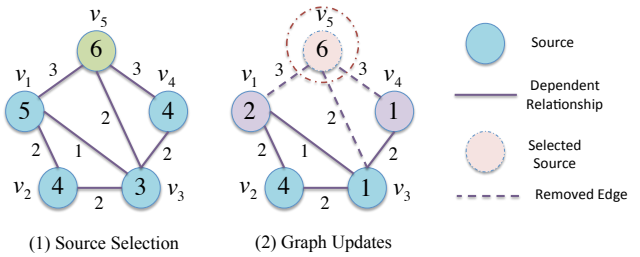

Figure 2. Simple Illustrative Example for CSS Scheme

In summary, the input to the CSS scheme is the generated graph G_s and the claims set C. The output of the CSS scheme is the set of critical sources S^*. The CSS scheme is summarized in Algorithm 1. The time complexity of the first step (i.e., vertex selection) is of order $O(|S|)$ and the time complexity of the second step (i.e., vertex weight update) is also of order $O(|S|)$. We iteratively conduct the above two steps until $C^*_{t+1} = C^*_t \cup C_{S_i}$. Therefore, the time complexity

Algorithm 1: Critical Source Selection (CSS) Scheme

Input: A weighted and undirected graph $G_s = (S, E_S, W_e, W_v)$ and the full set of claim C.
Output: A set of selected sources S^*.
Initialize: $S_0^* \leftarrow \emptyset$, $C_0^* \leftarrow \emptyset$, $G_s^0 \leftarrow G_s$, $E_S^0 \leftarrow E_S$, $t \leftarrow 0$
repeat
  Select the vertex in graph G_s with the largest vertex weight. (without loss of generality, we suppose that the selected node is v_i)
  for each $i' \in [1, ..., X]$ and $i' \neq i$ **do**
    $wv_{i'} \leftarrow wv_i - we_{i,i'} \cdot \varphi$
    $S_{t+1}^* \leftarrow S_t^* \cup \{v_i\}$
    $G_s^{t+1} \leftarrow G_s^t - \{v_i\}$
    $E_{t+1} \leftarrow E_t - \{e_{i,i'}\}$
    $C_{t+1}^* = C_t^* \cup C_{S_i}$
  end for
  $t = t + 1$
until $C_t^* = C$

of our CSS algorithm is of order $O(|S| \cdot |S^*|)$, where $|S^*|$ is the size of selected critical source set. Since $|S^*|$ is normally much smaller than $|S|$, the CSS scheme is scalalbe in large-scale social sensing applications. Finally, we can also prove our model is an additive model and the CSS scheme can find the optimal solutions under a normal condition that has been widely used in additive models [8].

V. Evaluation

In this section, we conduct experiments to evaluate the performance of the *CSS (Critical Source Selection)* scheme on two real-world data traces collected in a real world social sensing application. We demonstrate the effectiveness and efficiency of our proposed methods on these data traces and compare the performance of our scheme to the state-of-the-art baselines. In the rest of this section, 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 that *CSS* scheme can help to achieve better truth discovery results by judiciously selecting the critical set of sources.

A. Experimental Setups and Evaluation Metrics

1) Data Trace Statistics: In this paper, we evaluate our proposed scheme on two 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. On Twitter, users have both explicitly (e.g., following relationship) and implicit (e.g., retweet behavior) dependency and tweet with different speak rates. These features of Twitter users provide us a good opportunity to investigate the performance of the *CSS* scheme in real world social sensing scenarios. In the evaluation, we selected two data traces: (i) Paris Attack event that happened on Nov, 2015; (ii) Oregon Shooting that happened on Oct, 2015. These data traces were collected through Twitter open API using query terms and specified geographic regions related to the events. The statistics of the two data traces are summarized in Table II.

Table II
Data Traces Statistics

Data Trace	Paris Attack	Oregon Shooting
Start Date	11/13/2015	10/1/2015
Physical Location	Paris, France	Umpqua, Oregon
Search Keywords	Paris, Attacks, ISIS	Oregon, Shooting, Umpqua
# of Tweets	873,760	210,028

2) Data Pre-Processing: To evaluate our methods in real world settings, we went through the following data pre-processing steps to generate the inputs for the *CSS* scheme: (i) Source-Claim Matrix (i.e., SC Matrix); (ii) Speak-Rate Vector (i.e., SR Matrix); (iii) Source-Dependency-Score Matrix (i.e., SDS Matrix). They are summarized as follows:

- *Source-Claim Matrix Generation*: We generate the SC Matrix as follows: 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) [29]. 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. Second, we generate the SC Matrix by associating each source with the claims he/she reported. In particular, we set the element $SC_{i,j}$ in SC matrix to 1 if source S_i generates a tweet that belongs to claim (cluster) C_j and 0 otherwise.
- *Speak-Rate Vector Generation*: We generate the SR Vector based on the constructed SC Matrix from the previous step. In particular, element SR_i in SR is the number of claims reported by source S_i normalized by the total number of claims. Formally, $SR_i = \frac{\Sigma_{j=1}^{Y} SC_{i,j}}{Y}$.
- *Source-Dependency-Score Matrix Generation*: We generate the SDS Matrix based on source reporting behaviors on Twitter. In particular, we generated the source dependency graph as an arbitrary undirected graph $G_{sds} = (V_{sds}, E_{sds}, W_{sds})$ where V_{sds} represents sources, E_{sds} represents their dependency links,

W_{sds} represents their dependency degree. We used the following heuristic to generate the links in the graph G_{sds}: an undirected edge from source S_i to source $S_{i'}$ is added if there exists claim reported by both source $S_{i'}$ and S_i. We then constructed the Social-Dependency-Score Matrix SDS by setting the corresponding element $SDS_{i,i'}$ as the number of claims reported by both source S_i and $S_{i'}$. We note that the above heuristic is only first approximations to estimate source dependency 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 *CSS* scheme: *Precision*, *Recall*, *F1-measure* and *Accuracy*. Their definitions are given in Table III.

In Table III, 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.

Table III
METRIC DEFINITIONS

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

B. Evaluation of Our Methods

In this subsection, we evaluate the performance of the proposed *CSS* scheme and compare it to the state-of-the-art source selection techniques. The baselines we used include:

- *DS*: it selects a set of diversified sources by only considering the dependency between sources using a set of heuristic based approaches in social sensing applications [32].
- *FS*: it selects a set of sources by considering source freshness based on the source reporting behaviors [27].
- *PS*: it selects the subset of data sources to predict the state of all other sources by considering source dependency in order to minimize the prediction errors on disaster response [10].

To evaluate all source selection schemes, we use the selection results from different algorithms as input to the state-of-the-art truth discovery techniques that include:

- *IPSN16*: it explores topic relevance feature of claims and the arbitrary source dependency between sources to ascertain the correctness of claims [12].
- *Sensys15*: it solves the problem of truth discovery for correlated claims by modeling claims' correlations as regularization terms [23].
- *IPSN14*: it solves the truth discovery problem by explicitly modeling the dependency between sources on social networks using an estimation theoretic approach [34].

In our evaluation, we combine each source selection scheme with different truth discovery techniques on claim correctness estimation. We manually graded the output of these combinations to determine the correctness of the claims. Considering the manpower limitations, we took the union of the top 50 claims returned by different schemes as our evaluation set in order to avoid the bias towards any particular scheme. The following rubric is used to collect the ground truth information of the evaluation set:

- *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.

The evaluation results of Paris Attack data trace are shown in Table IV. We can observe that *CSS* scheme outperforms the compared baselines with different truth discovery techniques in all evaluation metrics. The largest performance gain achieved by *CSS* on F1-measure and accuracy over the best performed baseline (i.e., PS) are 10% and 9% respectively. The results of Oregon Shooting data trace are presented in Table V. We can observe that *CSS* scheme continues to outperform all baselines with different truth discovery techniques. The performance improvements of *CSS* are achieved by explicitly considering both the source dependency and source speak rate in sensor selection process, one of the main contributions of this paper.

Table IV
SOURCE SELECTION EVALUATION ON PARIS ATTACK DATA TRACE

Alg	Truth Discovery	Accuracy	Precision	Recall	F1-score
SS	IPSN16	**0.700**	**0.803**	**0.741**	**0.771**
	Sensys15	**0.637**	**0.704**	**0.825**	**0.760**
	IPSN14	**0.688**	**0.799**	**0.738**	**0.767**
DS	IPSN16	0.489	0.655	0.560	0.604
	Sensys15	0.546	0.700	0.610	0.652
	IPSN14	0.486	0.642	0.589	0.614
FS	IPSN16	0.567	0.788	0.516	0.624
	Sensys15	0.572	0.697	0.680	0.688
	IPSN14	0.617	0.787	0.618	0.692
PS	IPSN16	0.600	0.797	0.560	0.658
	Sensys15	0.572	0.697	0.680	0.689
	IPSN14	0.615	0.788	0.611	0.688

Table V
SOURCE SELECTION EVALUATION ON OREGON COLLEGE SHOOTING DATA TRACE

Alg	Truth Discovery	Accuracy	Precision	Recall	F1-score
SS	IPSN16	**0.648**	**0.702**	**0.803**	**0.749**
	Sensys15	**0.636**	0.685	**0.822**	**0.747**
	IPSN14	**0.633**	**0.696**	**0.780**	**0.735**
DS	IPSN16	0.522	0.681	0.509	0.583
	Sensys15	0.572	0.681	0.649	0.665
	IPSN14	0.519	0.683	0.495	0.574
FS	IPSN16	0.568	0.683	0.635	0.658
	Sensys15	0.581	**0.708**	0.612	0.656
	IPSN14	0.556	0.676	0.616	0.645
PS	IPSN16	0.559	0.678	0.621	0.649
	Sensys15	0.544	0.699	0.532	0.605
	IPSN14	0.544	0.673	0.588	0.628

VI. CONCLUSION

In this paper, we develop a new critical source selection in social sensing to effectively reduce the complexity of a truth discovery problem and improve the accuracy of estimation results at the same time. In particular, our proposed scheme (CSS scheme) explicitly explores the source dependency and speak rate in the solution of critical source selection. We perform extensive experiments to compare the performance of our CSS scheme with the-sate-of-the-art baselines using real-world social sensing datasets. The evaluation results demonstrate the effectiveness and efficiency achieved by our scheme.

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

[2] H. Amintoosi, S. S. Kanhere, and M. Allahbakhsh. Trust-based privacy-aware participant selection in social participatory sensing. *Journal of Information Security and Applications*, 20:11–25, 2015.

[3] M. Azizyan, I. Constandache, and R. Roy Choudhury. Surroundsense: mobile phone localization via ambience fingerprinting. In *Proceedings of International Conference on Mobile Computing and Networking (Mobicom)*, pages 261–272. ACM, 2009.

[4] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava. Participatory sensing. *Center for Embedded Network Sensing*, 2006.

[5] A. T. Campbell, S. B. Eisenman, N. D. Lane, E. Miluzzo, and R. A. Peterson. People-centric urban sensing. In *Proceedings of International Workshop on Wireless Internet*, WICON '06, New York, NY, USA, 2006. ACM.

[6] X. L. Dong, B. Saha, and D. Srivastava. Less is more: Selecting sources wisely for integration. In *Proceedings of the VLDB Endowment*, volume 6, pages 37–48. VLDB Endowment, 2012.

[7] A. Guéziec. Crowd sourced traffic reporting, May 6 2014. US Patent 8,718,910.

[8] T. J. Hastie and R. J. Tibshirani. *Generalized additive models*, volume 43. CRC Press, 1990.

[9] 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.

[10] M. Hosseini, N. Nagibolhosseini, A. Barnoy, P. Terlecky, H. Liu, S. Hu, S. Wang, T. Amin, L. Su, D. Wang, et al. Joint source selection and data extrapolation in social sensing for disaster response. *arXiv preprint arXiv:1512.00500*, 2015.

[11] C. Huang and D. Wang. Spatial-temporal aware truth finding in big data social sensing applications. In *Proceedings of Trustcom/BigDataSE/ISPA*, volume 2, pages 72–79. IEEE, 2015.

[12] C. Huang and D. Wang. Topic-aware social sensing with arbitrary source dependency graphs. In *International Conference on Information Processing in Sensor Networks (IPSN)*, pages 1–12. ACM/IEEE, 2016.

[13] C. Huang and D. Wang. Unsupervised interesting places discovery in location-based social sensing. In *Proceedings of International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 67–74. IEEE, 2016.

[14] C. Huang, D. Wang, and N. Chawla. Towards time-sensitive truth discovery in social sensing applications. In *Proceedings of International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, pages 154–162. IEEE, 2015.

[15] C. Huang, D. Wang, and N. Chawla. Scalable uncertainty-aware truth discovery in big data social sensing applications for cyber-physical systems. *IEEE Transactions on Big Data*, 2017.

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

[17] 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.

[18] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han. A confidence-aware approach for truth discovery on long-tail data. *VLDB Endowment*, pages 425–436, 2014.

[19] P. Marks. Crowds point out potholes on a map to speed up street repairs, 2013.

[20] J. Marshall, M. Syed, and D. Wang. Hardness-aware truth discovery in social sensing applications. In *Distributed Computing in Sensor Systems (DCOSS), 2016 International Conference on*, pages 143–152. IEEE, 2016.

[21] J. Marshall and D. Wang. Mood-sensitive truth discovery for reliable recommendation systems in social sensing. In *Proceedings of International Conference on Recommender Systems (Recsys)*, pages 167–174. ACM, 2016.

[22] J. Marshall and D. Wang. Towards emotional-aware truth discovery in social sensing applications. In *Smart Computing (SMARTCOMP), 2016 IEEE International Conference on*, pages 1–8. IEEE, 2016.

[23] C. Meng, W. Jiang, Y. Li, J. Gao, L. Su, H. Ding, and Y. Cheng. Truth discovery on crowd sensing of correlated entities. In *Proceedings of International Conference on Embedded Networked Sensor Systems (Sensys)*, pages 169–182. ACM, 2015.

[24] 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 International Conference on Information Processing in Sensor Networks (IPSN)*, pages 190–201. ACM/IEEE, 2015.

[25] R. Pochampally, A. Das Sarma, X. L. Dong, A. Meliou, and D. Srivastava. Fusing data with correlations. In *Proceedings of International Conference on Management of Data (SIGMOD)*, pages 433–444. ACM, 2014.

[26] S. Reddy, K. Shilton, G. Denisov, C. Cenizal, D. Estrin, and M. Srivastava. Biketastic: sensing and mapping for better biking. In *International Conference on Human Factors in Computing Systems (CHI)*, pages 1817–1820, New York, NY, USA, 2010. ACM.

[27] T. Rekatsinas, X. L. Dong, and D. Srivastava. Characterizing and selecting fresh data sources. In *Proceedings of International Conference on Special Interest Group on Management of Data (SIGMOD)*, pages 919–930. ACM, 2014.

[28] F. Restuccia, S. K. Das, and J. Payton. Incentive mechanisms for participatory sensing: Survey and research challenges. *Transactions on Sensor Networks (TOSN)*, 12, 2016.

[29] K. D. Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking. Topical clustering of tweets. In *Proceedings of International Conference on Special Interest Group on Information Retrieval (SIGIR)*. ACM, 2011.

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

[31] J. Teng, B. Zhang, X. Li, X. Bai, and D. Xuan. E-shadow: Lubricating social interaction using mobile phones. *Transactions on Computers (TOC)*, 63(6):1422–1433, 2014.

[32] M. Y. S. Uddin, M. T. Al Amin, H. Le, T. Abdelzaher, B. Szymanski, and T. Nguyen. On diversifying source selection in social sensing. In *Proceedings of International Conference on Networked Sensing Systems (INSS)*, pages 1–8. IEEE, 2012.

[33] D. Wang, M. Amin, T. Abedlzaher, D. Roth, C. Voss, L. Kaplan, S. Tratz, J. Laoudi, and D. Briesch. Provenance-assisted classification in social networks. *IEEE Journal of Selected Topics in Signal Processing (J-STSP)*, 2014.

[34] 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 *International Conference on Information Processing in Sensor Networks (IPSN)*, pages 35–46. ACM/IEEE, 2014.

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

[36] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher. On truth discovery in social sensing: A maximum likelihood estimation approach. In *International Conference on Information Processing in Sensor Networks (IPSN)*, April 2012.

[37] D. Wang, J. Marshall, and C. Huang. Theme-relevant truth discovery on twitter: An estimation theoretic approach. In *Proceedings of International Conference on Web and Social Media (ICWSM)*, pages 408–416. AAAI, 2016.

[38] L. Wang, D. Zhang, D. Yang, B. Y. Lim, and X. Ma. Differential location privacy for sparse mobile crowdsensing. In *Proceedings of International Conference on Data Mining (ICDM)*. IEEE, 2016.

[39] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su. Understanding retweeting behaviors in social networks. In *Proceedings of International on Conference on Information and Knowledge Management (CIKM)*, pages 1633–1636. ACM, 2010.

[40] 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.

[41] D. Y. Zhang, C. Zheng, D. Wang, D. Thain, C. Huang, X. Mu, and G. Madey. Towards scalable and dynamic social sensing using a distributed computing framework. In *The 37th IEEE International Conference on Distributed Computing (ICDCS 2017), in print*. IEEE, 2017.

[42] Z. Zhao, J. Cheng, and W. Ng. Truth discovery in data streams: A single-pass probabilistic approach. In *Proceedings of International Conference on Conference on Information and Knowledge Management (CIKM)*, pages 1589–1598. ACM, 2014.

[43] H.-J. Zimmermann. Fuzzy programming and linear programming with several objective functions. *Fuzzy sets and systems*, 1(1):45–55, 1978.