This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2824812, IEEE Transactions on Big Data

On Scalable and Robust Truth Discovery in Big Data Social Media Sensing Applications

Daniel (Yue) Zhang, Dong Wang, Nathan Vance, Yang Zhang, and Steven Mike

Abstract—Identifying trustworthy information in the presence of noisy data contributed by numerous unvetted sources from online social media (e.g., Twitter, Facebook, and Instagram) has been a crucial task in the era of big data. This task, referred to as truth discovery, targets at identifying the reliability of the sources and the truthfulness of claims they make without knowing either *a priori*. In this work, we identified three important challenges that have not been well addressed in the current truth discovery literature. The first one is "misinformation spread" where a significant number of sources are contributing to false claims, making the identification of truthful claims difficult. For example, on Twitter, rumors, scams, and influence bots are common examples of sources colluding, either intentionally or unintentionally, to spread misinformation and obscure the truth. The second challenge is "data sparsity" or the "long-tail phenomenon" where a majority of sources only contribute a small number of claims, providing insufficient evidence to determine those sources' trustworthiness. For example, in the Twitter datasets that we collected during real-world events, more than 90% of sources only contributed to a single claim. Third, many current solutions are not *scalable* to large-scale social sensing events because of the centralized nature of their truth discovery algorithms. In this paper, we develop a Scalable and Robust Truth Discovery (SRTD) scheme to address the above three challenges. In particular, the SRTD scheme jointly quantifies both the reliability of sources and the credibility of claims using a principled approach. We further develop a distributed framework to implement the proposed truth discovery scheme using Work Queue in an HTCondor system. The evaluation results on three real-world datasets show that the SRTD scheme significantly outperforms the state-of-the-art truth discovery methods in terms of both effectiveness and efficiency.

Index Terms—Big Data, Truth Discovery, Rumor Robust, Sparse Social Media Sensing, Scalable, Twitter

1 Introduction

This paper presents a new scalable and robust approach to solve the truth discovery problem in big data social media sensing applications. Online social media (e.g., Twitter, Facebook, and Instagram) provides a new sensing paradigm in the big data era where people act as ubiquitous, inexpensive, and versatile sensors to spontaneously report their observations (often called claims) about the physical world. This paradigm is motivated by the increasing popularity of portable data collection devices (e.g., smartphones) and the massive data dissemination opportunities enabled by online social media [31]. Examples of social media sensing include real-time situation awareness services in disaster or emergency response [37], intelligent transportation system applications using location-based social network services [35], and urban sensing applications using common citizens [49]. A critical challenge that exists in social media sensing is *truth discovery* where the goal is to identify reliable sources and truthful claims from massive noisy, unfiltered, and even conflicting social media data. The truth discovery problem stays in the heart of the *veracity* challenge of big data social media sensing applications.

To solve the truth discovery problem, a rich set of principled approaches have been proposed in machine learning, data mining, and network sensing communities [6], [7], [17], [22], [37], [51]. However, three important challenges have yet to be well addressed by existing truth discovery solutions in social media sensing applications.

- *The authors are with the Department of Computer Science and Engineering and the Interdisciplinary Center for Network Science and Applications (iCeNSA), University of Notre Dame, Notre Dame, IN 46556. E-mail: {yzhang40,dwang5,nvance1,yzhang42,smike}@nd.edu*

First, current truth discovery solutions do not fully address the "misinformation spread" problem where a significant number of sources are spreading false information on social media. For example, a piece of misinformation on Twitter saying that an 8-year-old girl was killed while running during the Boston Marathon has been so widely spread that the misinformation to debunking ratio was 44:1 [26]. In examples like this, the widely spread false information appears much more prominently than the truthful information, making truth discovery a challenging task. Our evaluation results on three real-world events demonstrate that current truth discovery solutions perform poorly in identifying truth when misinformation is widely spread. Second, many current truth discovery algorithms depend heavily on the accurate estimation of the reliability of sources, which often requires a reasonably dense dataset. However, "data sparsity" or the "long-tail phenomenon" [44] is commonly observed in real-world applications. For example, due to the spontaneous nature of social media sensing, sources might lack the motivation and incentives to continuously contribute data to the application [50]. Alternatively, sources might choose to ignore topics or events that they are not interested in and only contribute data to the topics or events that match their interests. In fact, in the real-world Twitter datasets we collected, over 90% of users only contribute a single tweet. In such a scenario where a vast majority of sources contribute only a small number of claims, there exists insufficient evidence for accurate estimation of source reliability. Li *et al.* and Xiao *et al.* have explicitly discussed the problem of data sparsity and demonstrated that many existing truth discovery algorithms fail to provide good estimations for source reliability when the dataset is sparse [7].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2824812, IEEE Transactions on Big Data

For example, in an extreme case where a user only posts one tweet, current truth discovery schemes will only be able to identify binary values of reliability (either 0 or 1), resulting in poor estimates of actual source reliability [44]. Third, existing truth discovery solutions did not fully explore the scalability aspect of the truth discovery problem [12]. Social sensing applications often generate large amounts of data during important events (e.g., disasters, sports, unrests) [23]. For example, during the 2016 Super Bowl, 3.8 million people generated a total of 16.9 million tweets with a peak rate of over 152,000 tweets per minute [19]. Current centralized truth discovery solutions are incapable of handling such a large volume of social sensing data due to the resource limitation of a single computing device. A few distributed solutions have been developed to address the scalability issue of the truth discovery problem [20]. However, they suffer from problems such as long startup times and ignorance of the heterogeneity of computational resources.

In this paper, we develop a Scalable and Robust Truth Discovery (SRTD) scheme to address the *misinformation spread*, *data sparsity*, and *scalability* challenges in big data social media sensing applications. To address the misinformation spread challenge, the SRTD scheme explicitly models various behaviors that sources exhibit such as copying/forwarding, self-correction, and spamming. To address data sparsity, the SRTD scheme employs a novel algorithm that estimates claim truthfulness from both the credibility analysis on the content of the claim and the historical contributions of sources who contribute to the claim. To address the scalability challenge, we develop a light-weight distributed framework using Work Queue [4] and HTCondor [27], which form a system that is shown to be both *scalable* and *efficient* in solving the truth discovery problem. We evaluate our SRTD scheme in comparison with state-of-the-art baselines on three real-world datasets collected from Twitter during recent events (Dallas Shooting in 2016, Charlie Hebdo Attack in 2015, and Boston Bombing in 2013). The evaluation results show that our SRTD scheme outperforms the state-of-the-art truth discovery schemes by accurately identifying the truthful information in the presence of widespread misinformation and sparse data, and significantly improving the computational efficiency.

We summarize our contributions as follows:

- We address three important challenges (i.e., misinformation spread, data sparsity, and scalability) in solving the truth discovery problem in big data social media sensing applications.
- We develop a novel Scalable Robust Truth Discovery (SRTD) scheme that explicitly considers various source behaviors, content analysis of claims, and historical contributions of sources in a holistic truth discovery solution.
- We develop a light-weight distributed framework based on Work Queue and HTCondor to implement the SRTD scheme and improve computational efficiency.
- We compare the performance of the SRTD scheme to a set of representative truth discovery solutions using three large-scale real-world datasets. The evaluation results demonstrate that the SRTD scheme achieves significant performance gains in terms of both effectiveness and efficiency compared to the baselines.

A preliminary version of this work has been published in [44]. We refer to the scheme developed in the previous work as the Reliable Truth Discovery (RTD) scheme. The current paper is a significant extension of the previous work in the following aspects. First, we extend our previous model by deriving the contribution score of sources using a more fine-grained and principled approach (Section 3). Specifically, we introduce the notion of Attitude Score, Uncertainty Score, and Independent Score to quantify the contributions from a source to the claim. The SRTD scheme is shown to be more accurate than the previous RTD scheme (Section 6). Second, we address the scalability challenge of the truth discovery solutions by developing a new distributed framework using Work Queue and HTCondor. We also implement a control system to optimize system performance (Section 4). Third, we add a new and more recent real-world dataset (i.e., Dallas Shooting in 2016) to further evaluate the performance and robustness of our proposed scheme in an additional real-world scenario (Section 6). Fourth, we compare our scheme with state-of-the-art baselines from recent truth discovery literature and demonstrate the performance improvements achieved by the SRTD scheme (Section 6). Finally, we extend the related work by reviewing recent works on the distributed truth discovery solutions (Section 2).

2 Related Work

2.1 Social Media Sensing

Social media sensing is an emerging sensing paradigm where social sensors (i.e. social media users) voluntarily report their observations of the physical world [38]. Combined with social media analysis techniques, social media sensing enables a great variety of applications. Examples include urban abnormality detection [5], social event summarization [3], user trajectory prediction [47], emergency response [39], and resource management [45]. This work focuses on the truth discovery problem where the goal is to jointly estimate the truthfulness of claims on social media and the reliability of social media users. The solution to this problem can benefit social media sensing applications by addressing the data veracity challenge in a noisy social media environment.

2.2 Truth Discovery

Truth discovery has received a significant amount of attention in recent years, and previous studies have developed various models to address this important challenge in big data applications. The truth discovery problem was first formally defined by Yin *et al.* [42], in which a Bayesian-based heuristic algorithm, *Truth Finder*, was proposed. Pasternack *et al.* extended this model by incorporating prior knowledge of constraints into truth discovery solutions and developed several solutions (*AvgLog*, *PooledInvest*) [22]. Dong *et al.* explicitly considered the source dependency in truth discovery problems [6]. A semi-supervised graph learning scheme was proposed to model the propagation of information truthfulness from the known ground truths [43]. Wang *et al.* proposed a scheme that offered a joint estimation on

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2824812, IEEE Transactions on Big Data

source reliability and claim correctness using maximum-likelihood estimation approach [37]. Zhang *et al.* developed a constraint-aware truth discovery model to incorporate physical constraints into detecting dynamically evolving truth [46]. However, there exists a significant knowledge gap in existing truth discovery solutions in terms of identifying truthful claims among widely spread misinformation, which is both a challenging and critical task in truth discovery. In our work, we propose a new truth discovery scheme that is robust against misinformation spread and is able to find truthful claims even if the majority of sources are providing misinformation.

2.3 Data Sparsity

Data sparsity or the "long-tail phenomenon" is an important challenge in many big data research areas [9], [18]. However, very few truth discovery schemes have explicitly considered this challenge even though sparse data is ubiquitous in real-world social media sensing applications. Li *et al.* proposed a Confidence-Aware Truth Discovery scheme (CATD) based on the observation that a point estimator for source reliability is not reliable when sources contribute very few claims. The CATD method derives a confidence interval to quantify the accuracy of source reliability estimation. Xiao *et al.* further extended the CATD model to explicitly consider the confidence interval of the truthfulness of the claims. They argued that when a claim has few sources contributing to it, the estimation score for the truthfulness of the claim becomes less meaningful. They proposed a new truth discovery scheme called Estimating Truth and Confidence Interval via Bootstrapping (ETCI-BooT) that was able to construct claims' confidence intervals as well as identifying the truth [7]. Although both of these works considered data sparsity, they did not evaluate their performance on detecting widespread misinformation. In fact, our evaluation results have suggested that the above solutions are not robust against widespread misinformation in social media sensing applications.

2.4 Distributed Systems for Social Sensing

Our work also bears some resemblance to a few distributed system implementations for social sensing applications. For example, Ouyang *et al.* developed a parallel algorithm for quantitative truth discovery applications to efficiently handle big streaming data by using the MapReduce framework in Hadoop [20]. Yerva *et al.* developed a cloud-serving system for fusing the social and sensor data to deal with massive data streams [41]. Xue *et al.* introduced a cloud-based system for large-scale social network analysis using the Hadoop framework [40]. A limitation of these approaches is that Hadoop is designed for dealing with large datasets and is too heavy-weight for time-critical applications that require fast response times in the presence of both small and large datasets [20]. In this work, we develop a light-weight distributed framework using Work Queue and HTCondor to improve the efficiency of our truth discovery scheme. This framework is ideal for time-critical systems because i) HTCondor is a high throughput distributed computing system that allows parallel computation of thousands of tasks, thus significantly reducing the overall processing time [48]; ii) the flexible priority scheduling allows critical tasks to be processed faster to meet the deadline requirements [28]; iii) the initialization time of HTCondor jobs compared to Hadoop is much smaller, making it more suitable to handle streaming data.

3 Problem Formulation

In this section, we formulate our robust truth discovery problem in big data social media sensing. In particular, consider a social media sensing application where a group of M sources $S = (S_1, S_2, ..., S_M)$ reports a set of N claims, namely, $C = (C_1, C_2, ..., C_N)$. Let S_i denote the ith source and C_j denote the jth claim. We define $RP^t_{i,j}$ to be the report made by source S_i on claim C_j at time t.

Take Twitter as an example; a source refers to a user account and a claim is a statement of an event, object, or topic that is derived from the source's tweet. For example, a tweet "Not much of the comment about the Dallas shooting has focused on the fact the sniper was a veteran." is associated with a claim "Dallas shooting sniper was a veteran". The tweet itself is considered as the report. We observe that the social media sensing data is often sparse (i.e., the majority of sources only contribute to a limited number of claims in an event).

We further define $C_j = T$ and $C_j = F$ to represent that a claim is true or false, respectively. Each claim is also associated with a ground truth label $\{x^*_j\}$ such that $x_j = 1$ when C_j is true and $x_j = 0$ otherwise.

The goal of the truth discovery task is to jointly estimate the truthfulness of each claim and the reliability of each source, which is defined as follows:

DEFINITION 1. **Claim Truthfulness D_j for claim C_j**: The likelihood of a claim to be true. The higher D_j is, the more likely the claim C_j is true. Formally we define D_j to estimate:

$$Pr(C_j = T) \tag{1}$$

DEFINITION 2. **Source Reliability R_i for source S_i**: A score represents how trustworthy a source is. The higher R_i is, the more likely the source S_i will provide credible and trustworthy information. Formally we define R_i to estimate:

$$Pr(C_j = T | SC_{i,j} = T) \tag{2}$$

where $SC_{i,j} = T$ denotes that source S_i reports claim C_j to be true.

Since sources are often unvetted in social media sensing applications and may not always report truthful claims, we need to explicitly model the reliability of data sources in our problem formulation. However, it is challenging to accurately estimate the reliability of sources when the social media sensing data is sparse [34]. Fortunately, the reports themselves often contain extra evidence and information to infer the truthfulness of a claim. In the Twitter example, the text, pictures, URL links, and geotags contained in the tweet can all be considered as extra evidence of the report. To leverage such evidence in our model, we define a *credibility score* for each report to represent how much the report contributes to the truthfulness of a claim.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2824812, IEEE Transactions on Big Data

We first define the following terms related to the credibility score of a report made by source S_i on claim C_j at time k.

DEFINITION 3. **Attitude Score ($\rho_{i,j}^k$)** : Whether a source believes the claim is true, false or does not provide any report. We use 1, -1 and 0 to represent these attitudes respectively.

DEFINITION 4. **Uncertainty Score ($\kappa_{i,j}^k$)**: A score in the range of (0,1) that measures the uncertainty of a report. A higher score is assigned to a report that expresses more uncertainty.

DEFINITION 5. **Independent Score: ($\eta_{i,j}^k$)**: A score in the range of (0,1) that measures whether the report $R_{i,u}$ is made independently or copied from other sources. A higher score is assigned to a report that is more likely to be made independently.

Combining the above terms, we formally define the *Credibility Score* of a report from source S_i on claim C_j at time k as:

$$SLS_{i,j}^k = \rho_{i,j}^k \times (1 - \kappa_{i,j}^k) \times \eta_{i,j}^k \tag{3}$$

In Equation (3), we make the assumption that the credibility of a report depends on a set of semantic scores related with the report, namely, attitude score, uncertainty score, and independence score. Using the above definition, we can clearly differentiate the reports on a claim in the following dimensions: i) a report that agrees or disagrees with the claim; ii) a report made with high or low confidence on the claim; iii) an original, copied, or forwarded report on the claim. All these factors are shown to be important in identifying truthful claims from widely spread misinformation [44].

Our model also explicitly considers a source's historical reports on the same claim. For example, spammers on Twitter can keep posting the exact same tweets over and over, which in most cases contain either irrelevant or misleading claims. On the other hand, a reliable source such as a police department or a responsible news outlet may proactively correct its previous reports that carry misinformation. Therefore, we define a time-series matrix to explicitly model the historical contributions of a source on its claims.

Given M sources and N claims, we define a *Time-series Source Claim (TSC)* matrix $TSC_{M\times N}$ where each element $\{SLS_{i,j}^k\}$ represents the historical credibility score of a report from source S_i on claim C_j at the time instance k.

$$TSC_{ij} = \{SLS_{i,j}^1, SLS_{i,j}^2, \cdots SLS_{i,j}^k, \cdots\} \tag{4}$$

The defined parameters and variables are summarized in Table 1. Using the above definitions, we can formally define the robust truth discovery problem in big data social media sensing applications as follows: given the Time-series Source-Claim Matrix TSC generated from the social media sensing data as input, the objective is to estimate the truthfulness D_j of each claim as the output. Specifically, we compute:

$$\forall j, 1 \leq j \leq N : Pr(C_j = T|TSC) \tag{5}$$

Table 1
Definition and Notation

S_i	The ith source
C_j	The jth claim
R_i	The reliability of the ith source
D_j	The truthfulness of the jth claim
$SLS_{i,j}^k$	The kth credibility score of the report from S_i on C_j
x_j^*	The ground truth label of the jth claim.
$\hat{x}_j^*$	Estimated label of the jth claim.

4 SOLUTION

In this section, we present the Scalable Robust Truth Discovery (SRTD) scheme to solve the truth discovery problem in big data social media sensing applications formulated in the previous section. We first outline a few observations relevant to our model. We then discuss the design intuition and present the SRTD scheme.

4.1 Observations

We find the following observations to be relevant to our model:

- Observation 1: Sources often spread false information by simply copying or forwarding information from others without independent verification (e.g., retweets on Twitter).
- Observation 2: False claims are often controversial and sources tend to disagree with each other and have intensive debates on those claims.
- Observation 3: If a source debunks its previous claim, it's very likely the previous claim is false because people are generally prone to be self-consistent.

4.2 Algorithm Design

Before delving into the details of the proposed SRTD scheme, we briefly review the current landscape of the truth discovery solutions in social media sensing. The current truth discovery solutions can be mainly classified into two categories: (i) principled solutions where explicit objective functions are defined and specific optimization techniques are used to find the convergence points at the local/global optimum of the objective functions (e.g., MLE, MAP based solutions [33], [34], [36], [37], [51]); (ii) data-driven solutions where heuristic based techniques (e.g., HITS, TruthFinder, AvgLog [14], [22], [42]) are adopted to address some practical data driven challenges (e.g., data sparsity) that are not well addressed by the principled solutions.

We observe that the principled solutions often work well on relatively dense datasets (e.g., the number of claims reported per source is high) but fail in the sparse data scenarios. The main reason is that the results of the principled solutions primarily depend on the accuracy of a potentially large set of estimation parameters (e.g., the parameters related to the source reliability and claim truthfulness), which are sensitive to the density of the observed data [9], [18]. In contrast, the data-driven solutions are often more heuristic by nature and explore the content of the sensing data to compensate for the data sparsity problem [22], [44].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2824812, IEEE Transactions on Big Data

Our SRTD scheme in this paper belongs to the category of data-driven solutions. It follows the intuition of our previous work [44] where the semantics of the tweets are found to be crucial in determining the claim truthfulness when the source reliability is hard to estimate given the sparse data. We compared SRTD with a few state-of-the-art principled truth discovery schemes (e.g., EM-SD [36], EM-Conflict [34]) in our evaluations (Section 6) and found that SRTD significantly outperformed those baselines when the data is sparse. Finally, we also discuss the future work of developing principled and robust truth discovery solutions for sparse social media sensing in Section 7.

4.3 Contribution Score (CS) of Sources

In the SRTD scheme, we first introduce the concept of a *Contribution Score (CS)* to quantify the actual contribution of a source on a claim. Using the TSC matrix defined in the previous section, we aggregate the credibility scores of all historic reports made by a source and define the *Contribution Score* of the source as follows:

DEFINITION 6. **Contribution Score** CS_{ij}: The source S_i's aggregated contribution to claim C_j, which is a function of the source reliability and the credibility scores of all historical reports made by the source.

In particular, the contribution score is calculated using the following rubrics:

- A more reliable source should be assigned a higher contribution score.
- Original reports of a claim should be assigned higher contribution scores than simply copying and forwarding reports.
- Reports with more assertion (i.e., less uncertainty) should be assigned higher contribution scores than those that express uncertainty or guesses.
- The self-correction behavior represents the reflection ability of the source which should be honored by assigning a higher contribution score to the source.
- Spamming behavior (i.e. a source keeps on forwarding the same claim) should be punished by decreasing the contribution score.

More specifically, the contribution score of source S_i on claim C_j is denoted as CS_{ij} and it is formally calculated as:

$$CS_{ij} = sgn(SLS^K_{i,j}) \sum_{k=1}^{K} R_i^{K+1-k} |SLS^k_{i,j}| \tag{6}$$

where R_i denotes the reliability of source S_i, $SLS^k_{i,j}$ denotes S_i's historical credibility score of a report made at time k on claim C_j, $sgn(SLS^K_{i,j})$ represents the sign of $SLS^K_{i,j}$, and K denotes the size of $SLS_{i,j}$ sequence. Since the sign of the formula only depends on the latest report, we honor the "self-correction" behavior by treating only the source's last report as its actual attitude towards the claim. We use term R_i^{K+1-k} as a "damping factor" to assign higher weights to "fresher" reports. The benefits are twofold: i) we reduce the effect of spamming behavior: if a user keeps tweeting the same thing over time, the old spamming reports will have little effect on the global contribution score of the user; ii) we assign the highest weight to the latest report from a source that debunks its own previous claims to alleviate the influence of their previous "mistakes". The definition of credibility scores also allows us to punish copying behaviors and unsure conjectures by assigning them lower scores as explained in the previous section.

4.4 SRTD Algorithm

The SRTD algorithm is an iterative algorithm that jointly computes the claim truthfulness and source reliability by explicitly considering the contribution scores of the sources. We initialize the model with uniform claim truthfulness scores and uniform source reliability scores. In each iteration, we first update the reliability score of each source using the truthfulness scores of claims reported by the source as well as the contribution score of the source itself. In particular, we compute the source reliability R_i of S_i as follows:

$$R_i = \frac{\sum_{j \in F(i)} |CS_{ij}| (\chi(CS_{ij}) D_j + (1 - \chi(CS_{ij}))(1 - D_j))}{\sum_{j \in F(i)} |CS_{ij}|}$$

$$\chi(a) = \begin{cases} 1, a > 0 \\ 0, a \leq 0 \end{cases} \tag{7}$$

where $F(i)$ is the set of the claims reported by source S_i. The above equation follows the intuition that the reliability of a source is proportional to the percentage of truthful claims it has provided. In particular, we consider the source's actual contribution to a claim by exploring its credibility score, which represents the source's attitude, uncertainty, and independence.

The next step of the iteration is to update the claim truthfulness score based on the newly computed source reliability. In particular, the truthfulness D_j of a claim C_j is calculated as:

$$\begin{aligned} TC_j &= \sum_{i \in K(j)} CS_{ij} \\ &= \sum_{i \in K(j)} sgn(SLS^K_{i,j}) \sum_{k=1}^{K} R_i^{K+1-k} |SLS^k_{i,j}| \quad (8) \\ D_j &= \frac{1}{1 + exp(-TC_j)} \quad (9) \end{aligned}$$

where $K(j)$ denotes all the sources who contributed to claim C_j. The above equation follows the intuition that a claim is more likely to be true if many reliable sources provide independent statements that assert the claim to be true. Unlike the previous models that only consider source reliability in the computation of the claim truthfulness [22], [42], our model explicitly incorporates both the historical contributions of a source and the credibility scores of reports made by the source.

In order to make the algorithm scalable, we first divide the input matrix TSC into a set of Z submatrices $TSC_1, TSC_2, ..., TSC_Z$ where each submatrix contains a subset of the sources in S and all of the claims that are reported by that subset of sources. In particular, we divide the TSC matrix such that all resulting submatrices contain a similar number of sources, and the size of each submatrix

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2824812, IEEE Transactions on Big Data

can easily fit into the memory of a single machine. It is worth noting that we do not require the submatrices to be of exactly the same size in the implementation. We designed and implemented a dynamic tuning scheme to effectively allocate submatrices to nodes in a cluster (discussed in details in the next section). We also observe that it makes sense to split the data into subsets of similar sizes in the first place in order to avoid unnecessary system tuning overhead that may occur later.

For each submatrix, we derive the source reliability and claim truthfulness scores independently, and then aggregate the results before the next iteration. Figure 1 shows a simple illustrative example of the algorithm. In this example, we have 6 sources and 5 claims. The algorithm first divides the data into three submatrices with 2 sources in each submatrix. For each iteration, we first compute the source reliability scores within each submatrix, which is an independent operation because it contains all claims made by its sources. Meanwhile, we compute the partial claim truthfulness scores (namely $< TC_1^z, TC_2^z ... TC_5^z >$) of all claims from the reliability scores of sources within each submatrix. The partial truthfulness scores of claims are then aggregated across the submatrices to obtain the final claim truthfulness scores. We use a sigmoid function to normalize the claim truthfulness scores. We finally update the contribution scores based on R_i and D_j before entering the next iteration of the SRTD algorithm. We note that the aggregation step requires exchanging the results of each submatrix in order to calculate the truthfulness scores of the claims. In our implementation, we use a shared directory to store, share, and update intermediate results in the SRTD scheme. The details of our implementation are discussed in the next section.

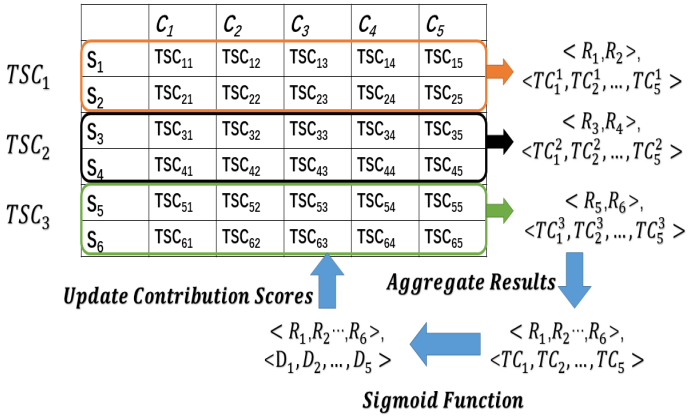


Figure 1. Dividing TSC Matrix Example with M = 6, N = 5, Z=3

Since each submatrix contains all of the claims that its sources contributed to, the source reliability scores can be calculated using the truthfulness of claims contributed by the source within the submatrix based on Equation (7). To calculate the truthfulness of the jth claim D_j, we first sum over contribution scores from all the sources who contribute to C_j within each submatrix TSC_z, denoted as TC_j^z. In particular, we calculate:

$$TC_j^z = \sum_{i \in K(j,z)} CS_{ij} \quad (10)$$

where $K(j,z)$ is the set of the sources who contribute to the jth claim in the z-th submatrix. We then derive the claim truthfulness scores as follows.

$$TC_j = \sum_{1 \le z \le Z} TC_j^z \quad (11)$$

$$D_j = \frac{1}{1 + exp(-TC_j)} \quad (12)$$

The pseudocode of the SRTD scheme is summarized in Algorithm 1.

Algorithm 1 Scalable Robust Truth Discovery (SRTD)

Input: TSC matrix
Output: claim truthfulness $\hat{x}_j^*, \forall 1 \le j \le N$
Initialize $R_i = 0.5, \forall i \le M$; set the values of credibility scores; initialize max_iteration = 100
Split Original TSC matrix into Z submatrices, let S(z) denote the number of sources in the z-th submatrix
while $\{D_j\}$ do not converge or reach max_iteration **do**
  for all $z, 1 \le z \le Z$ **do**
    for all $i, 1 \le i \le S(z)$ **do**
      for all $j, 1 \le j \le N$ **do**
        if TSC_{ij} exists **then**
          compute CS_{ij} based on Equation (6)
        end if
      end for
    end for
    for all $i, 1 \le i \le S(z)$ **do**
      estimate R_i based on Equation (7)
    end for
    for all $j, 1 \le j \le N$ **do**
      compute TC_j^z based on Equation (10)
    end for
    estimate D_j based on Equations (11) and (12)
  end for
end while
for all $j, 1 \le j \le N$ **do**
  if $D_j \ge threshold$ **then**
    output $\hat{x}_j^* = 1$
  else
    output $\hat{x}_j^* = 0$
  end if
end for

Initially, the algorithm has little information about the reliability of sources and the truthfulness of claims. In each iteration, SRTD improves its knowledge by jointly updating reliability scores and truthfulness scores until stopping criteria are met. The algorithm stops when it converges (i.e. the inter-iteration difference of claim truthfulness score is negligible) or a maximum number of iterations is reached.

4.5 Convergence and Complexity Analysis

Our algorithm adopts an iterative estimation method similar to TruthFinder [42], AverageLog [22], and RTD [44]. The results from previous works show quick convergence and meaningful results obtained by this iterative method [44]. In our work, we explicitly study the effectiveness and convergence of the SRTD scheme in Sections 6.2.4 and 6.2.6, respectively. The results show effective truth discovery results and quick convergence of the proposed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2824812, IEEE Transactions on Big Data

scheme over real-world social sensing data traces. The overall complexity of the SRTD scheme is $O(MNK)$ where M is the number of sources, N is the number of claims, and K is the number of iterations. In particular, the complexity of computing contribution scores $CS_{i,j}$, reliability scores (R_i), and truthfulness scores (D_j) is $O(MN)$, $O(N)$, and $O(N)$ respectively. Given the quick convergence of the SRTD scheme (i.e., Figure 10 of Section 6.2.6), K is often a small constant. Therefore, the overall complexity of the algorithm can be further simplified as $O(MN)$. Additionally, SRTD is a distributed solution that spreads the computation tasks over multiple nodes running in parallel, further improving the efficiency compared to centralized solutions (see Table 3 in Section 6.2.5 for details).

5 Implementation

In this section, we present a distributed implementation of the SRTD system using HTCondor and Work Queue. We first introduce the HTCondor system and Work Queue framework. Then, we present the implementation of the SRTD scheme, focusing on the allocation, management, and control of the distributed truth discovery tasks.

5.1 HTCondor and Work Queue

5.1.1 HTCondor

We use HTCondor at the University of Notre Dame as the underlying distributed system for the implementation of our SRTD scheme. The system consists of over 1,900 machines and over 14,700 cores at the time of writing. HTCondor has been used by hundreds of organizations in industry, government, and academia to manage computing clusters ranging from a handful to many thousands of workstation cores [16]. The HTCondor system at the University of Notre Dame is deployed to library desktops, server clusters, and other available machines which would otherwise be left unused 90% of the day. Users can submit their computation tasks to the HTCondor system, and the system allocates the tasks to run on idle machines that are connected to the system.

5.1.2 Work Queue

Work Queue is a lightweight framework for implementing large-scale distributed systems [4]. This framework allows the master process to define a set of tasks (i.e., Task Pool), submit them to the queue, and wait for completion. Work Queue maintains an elastic worker pool that allows users to scale the number of workers up or down as required by their applications. A worker is defined as a process that performs specific computational functions described by the tasks. Once running, each worker calls back to the master process, arranges for data transfers, and executes the tasks. We use Work Queue on top of the HTCondor system to take advantage of its dynamic resource allocation mechanism for task allocations.

5.2 Overview of SRTD Architecture

The architecture of the implemented SRTD system is shown in Figure 2. A key component is the Dynamic Task Manager (DTM), which is implemented as a master Work Queue process that initializes a Worker Pool and dynamically spawns new tasks into the Task Pool. The DTM first divides the original TSC matrix into submatrices as described in the previous section. Then, it spawns a set of tasks to process all submatrices in parallel on the HTCondor system. A feedback control system is integrated with the SRTD scheme to monitor the current execution speed of each Truth Discovery (TD) task and estimate its expected finish time. The feedback control system informs the DTM of control signals based on system performance, and it dynamically adjusts the task priority and resource allocation to optimize the overall system performance.

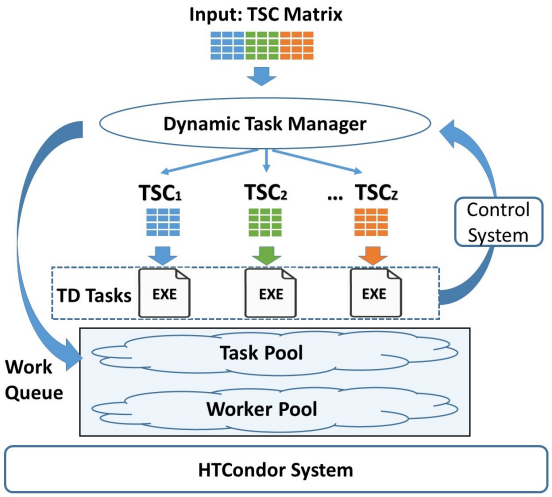


Figure 2. SRTD System Overview

5.3 Distributed Task Allocation

To make SRTD a scalable scheme, we divide the input data into multiple subsets and process them in parallel. In particular, we first divide the TSC matrix into Z submatrices $TSC_1, TSC_2, ..., TSC_Z$. The DTM then kicks off a TD task for each submatrix. The TD task performs the following operations:

1) Compute the source reliability based on equation (7)
2) Compute partial claim truthfulness TC_j^z based on equation (10)
3) Wait for all other TD tasks related to C_j to finish, and compute D_j by aggregating all partial claim truthfulnesses of C_j based on equations (11) and (12)
4) Update the Contribution Scores of sources based on equation (6)
5) Repeat the above steps until SRTD converges

Note that the third step requires sharing information among different TD tasks. We achieve this by sharing a common directory between TD tasks in the HTCondor system [1]. After computing the partial claim truthfulness and source reliability scores, each TD task records the intermediate results (i.e. TC_j^z and R_i) into a file in the shared directory. Before the end of each iteration, the DTM aggregates the

1. The HTCondor system does not provide a direct way to share information between different tasks.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2824812, IEEE Transactions on Big Data

results from files and updates the submatrices for each TD task. We note that this sharing mechanism introduces I/O overhead to the performance of the SRTD scheme. However, we found that this I/O overhead is relatively small compared to the total execution time of the truth discovery algorithm, which is shown in the evaluation results in the next section.

5.4 Dynamic Task Management and Feedback Control System

In this subsection, we discuss the implementation details of the Dynamic Task Manager (DTM). The DTM is designed to dynamically manage the tasks and resources to optimize the system performance. SRTD is essentially an iterative algorithm and has to wait for all of the TD tasks in the current iteration to finish before starting the next. Therefore, it is crucial to ensure that each TD task is synchronized and runs at the approximately same speed. We designed and implemented a feedback control system that allows us to monitor and dynamically control the speed of TD tasks.

The architecture of the dynamic feedback control system is shown in Figure 3. It consists of three key components: a Feedback Controller, Local Control Knobs (LCK), and a Global Control Knob (GCK). A LCK refers to a local control variable that is used to tune the performance of a particular TD task. In the SRTD scheme, a LCK is the subtask spawning of each TD task. Specifically, the LCK can split one TD task into multiple sub-tasks that run in parallel to improve the execution time of the TD task. The GCK refers to a global control variable that is used to tune the performance of all TD tasks in the system. In the SRTD scheme, we use the total number of workers in the Worker Pool as the GCK. In particular, we can change the resources that are assigned to all TD tasks by scaling up or down the number of workers in the pool.

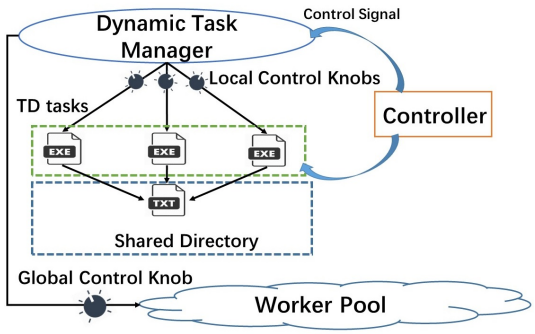


Figure 3. Dynamic Feedback Control System

The feedback control mechanism is designed to identify the lagging TD tasks. In particular, we monitor the execution time for each TD task at each iteration. The control signal is generated for a task that runs significantly slower than others and the signal is sent to DTM. After each iteration, the DTM tunes the control knobs using the following rubrics:

- *LCK Tuning:* If the i-th TD task is significantly slower than other tasks, the DTM splits the TSC_i into two submatrices with a similar number of sources. The DTM then spawns two sub-TD tasks for each matrix. The original TD task is removed from the task pool.
- *GCK Tuning:* If the execution times of all TD tasks are similar, and the execution time of the slowest task is lower than a certain threshold, we scale up the total number of workers by a factor of α.
- Otherwise, we enter the next iteration without system tuning.

6 Evaluation

In this section, we evaluate the SRTD scheme in comparison with the state-of-the-art truth discovery schemes on three real-world datasets collected from Twitter in recent events. The results demonstrate that the SRTD scheme significantly outperforms all of the compared baselines in terms of both accuracy and efficiency in big data social media sensing applications.

6.1 Experimental Setups

6.1.1 Baseline Methods

We chose 9 representative truth discovery solutions as the baselines in the evaluation: AvgLog, Invest, TruthFinder, 2-Estimates, CATD, ETBoot, EM-SD, EM-Conflict, and RTD.

- **AvgLog:** AvgLog is a basic truth discovery algorithm that jointly estimates the source reliability and claim truthfulness using a simple average-log function [22].
- **Invest:** The invest algorithm "invests" a source's reliability score among its claims. The truthfulness of each claim can be obtained using a nonlinear function [22].
- **TruthFinder:** TruthFinder identities trustworthy data sources and true facts by utilizing the interdependence between source trustworthiness and fact confidence [42].
- **2-Estimates:** 2-Estimates algorithm identifies the truthfulness of each claim by estimating the two defined parameters in their models related to source reliability and claim truthfulness [10].
- **CATD, ETBoot:** These two methods provide interval estimators for source reliability in a sparse dataset where most sources make one or two claims [7].
- **EM-SD, EM-Conflict:** These two methods both use a maximum likelihood estimation approach for truth discovery with social sensing data. EM-SD explicitly considers the retweeting behavior of Twitter users and uses that to build a source dependent model [36]. EM-Conflict explicitly considers conflicting claims from different sources [34].
- **RTD:** RTD is a robust truth discovery solution we developed in our previous conference paper. It leverages the semantic scores of tweets to identify the misinformation spread in social media sensing applications [44].

6.1.2 Parameter Tuning

Our scheme only has two parameters, the initial source reliability scores - R_i and the output threshold as described in Algorithm 1. We set them both to 0.5. For a fair comparison, we chose the parameter assignment that yields the best result for baselines that have a tunable parameter.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2824812, IEEE Transactions on Big Data

6.2 Experiment on Real World Data

6.2.1 Data Collection

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 [2]. The first one is the Dallas Shooting that happened on July 7, 2016. A heavily armed army veteran ambushed a group of police officers in Dallas, killing five officers and injuring nine others. The shooting was believed to be the deadliest incident for U.S. law enforcement since the 9/11 attacks. The second trace is the Paris Charlie Hebdo Attack which happened on Jan. 1, 2015. Several gunmen attacked the offices of a French satirical news magazine in Paris, killing 12 people including employees and two police officers. The third data trace is the Boston Bombing that happened on April 15, 2013, when two bombs were detonated near the finish line of the annual Boston Marathon, causing three deaths and injuring several hundred others.

We developed a data crawler based on the Twitter open search API to collect these data traces by specifying query terms and the geographic regions related to the events. The statistics of the three data traces are summarized in Table 2. We noted that all three datasets are very sparse. In the Dallas Shooting dataset, only 1.4% of sources contribute more than two claims while 91.5% of sources contribute only one claim. Similarly, in the Charlie Hebdo dataset, 90.8% of sources make a single claim and only 2.3% of sources provide more than two claims. In the Boston dataset, 94.9% of sources only make a single claim and only 1.1% of sources contribute more than two claims.

Our empirical results have shown that a large portion of tweets is related to misinformation. In particular, in our dataset, we found the proportions that are related to misinformation (the number of tweets that contribute to false claims) are 24.17%, 22.16% and 31.34% for the Dallas Shooting, Charlie Hebdo Shooting and Boston Bombing, respectively. Also, a vast majority of these misinformation related tweets (79.15%, 82.33%, 81.40% for the Dallas Shooting, Charlie Hebdo Shooting and Boston Bombing, respectively) are simply retweets or copies.

6.2.2 Data Preprocessing

We conducted the following data preprocessing steps to prepare the datasets for the experiment: (i) cluster similar tweets into the same cluster to generate claims; (ii) derive semantic link scores; (iii) generate the TSC Matrix; and (iv) generate ground truth labels. The details of these steps are summarized below.

Clustering: we first grouped similar tweets into the same cluster using a variant of K-means algorithm that can effectively handle streaming Twitter data [13] and the Jaccard distance to calculate the "distance" (i.e., similarity) between tweets [29]. This metric is very commonly used in clustering microblog posts (e.g., tweets) and has been shown to be effective in identifying similar tweets in truth discovery solutions [25], [44]. For each generated cluster, we picked a representative statement as the claim and we take each Twitter user as the source for our model described in Section 3.

Computing Credibility Score: To compute the Credibility Score of a report (i.e., tweet), we first calculated the Attitude Score of a source by performing a combination of sentiment analysis and keyword matching. Specifically, we performed Polarity Analysis to detect tweets that express strong negative sentiment (with negativity value ≤ 0.6) as "disagree" using the Python NLTK toolkit [3]. We further captured disagreeing tweets based on whether it contains certain keywords such as "fake", "false", "debunked", "rumor", "wrong", and "not true". We assigned a score of "1" and "-1" for non-disagreeing and disagreeing tweets respectively. We then calculated the *Uncertainty Score* by implementing a simple text classifier using skit-learn and trained it with the data provided by CoNLL-2010 Shared Task [8]. To compute the *Independent Score*, we implemented a script to automatically label a tweet as "dependent" if it is i) a retweet; or ii) significantly similar to other tweets (i.e., with a Jaccard distance less than 0.1) that were posted with earlier timestamps. We assigned a relatively low score to these dependent tweets. To evaluate the effectiveness of the above methods, we randomly picked 200 labeled tweets and manually checked the derived scores. The classification accuracy for *Attitude Score*, *Uncertainty Score* and *Independent Score* are 82.1%, 78.6%, and 90.3% respectively.

Generating the Time-series Source-Claim Matrix: We generated the TSC matrix as follows: for source S_i, we recorded all of its reports (i.e., tweets) that were related to C_j based on the clustering results. We sorted the tweets in chronological order and for each tweet, we derived the credibility score based on equation (3). The resulting time-series vector $\{SLS_{i,j}^1, SLS_{i,j}^2, \cdots SLS_{i,j}^k\}$ was stored as the element TSC_{ij} in the matrix.

Labeling Ground Truth: We first manually checked if a claim was a Factual Claim or an Unconfirmed Claim based on the following rubric:

- *Factual Claims*: Claims that are statements of a physical or social event related to the selected topic (i.e., Dallas Shooting, Charlie Hebdo Attack, or Boston Bombing) and observable by multiple independent observers.
- *Unconfirmed Claims*: Claims that do not meet the above criteria. Examples include the tweets that represent personal feelings, shout-outs, quotes from lyrics or books, etc.

After the above label processing, we discarded the unconfirmed claims (since they cannot be independently verified) and manually verified the truthfulness of the factual claims based on historical facts from credible sources external to Twitter (e.g., mainstream news media).

6.2.3 Evaluation Metrics

After labeling the ground truth of the claims, we noticed that our evaluation datasets are imbalanced: there were many more true claims than false claims. For example, only 19% of the claims in the Dallas Shooting data trace, 14% in the Charlie Hebdo Attack data trace, and 23% in the Boston Bombing data trace are false. However, this does not necessarily indicate that our truth discovery problem

2. http://apollo.cse.nd.edu/datasets.html

3. http://text-processing.com/demo/sentiment/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2824812, IEEE Transactions on Big Data

Table 2
Data Trace Statistics

Data Trace	Dallas Shooting	Charlie Hebdo Attack	Boston Bombing
Start Date	July 7 2016	January 1 2015	April 15 2013
Time Duration	8 days	3 days	4 days
Location	Dallas, USA	Paris, France	Boston, USA
Search Keywords	Dallas, Shooting, Police, Sniper	Paris, Shooting, Charlie Hebdo	Bombing, Marathon, Attack
# of Tweets (Original Set)	128,483	253,536	513,609
# of Tweets (Evaluation Set)	48,197	60,559	73,331
# of Tweets per User	1.17	1.09	1.14
% of Tweets Related To Misinformation	24.17%	22.86%	31.34%

on these data traces is simple; the false claims share many similar features as the true claims in the evaluation datasets (e.g., they have a similar number of supporting tweets, they are reported by similar sets of sources, etc.). Misclassifying a false claim as true can lead to significant negative impact in emergent events like disasters. To handle the data imbalance, we chose *Specificity (SPC)*, *Matthews Correlation Coefficient (MCC)*, and *Cohen's Kappa (Kappa)* [44] for imbalanced classification to evaluate the effectiveness of all compared schemes.

6.2.4 Evaluation Results - Truth Discovery Effectiveness

In the first set of experiments, we evaluate the effectiveness of truth discovery using above-mentioned metrics (i.e. SPC, MCC and Kappa). The results of the Dallas Shooting dataset are shown in Figure 4. We can observe that the SRTD scheme outperforms all of the compared baselines. Compared to the best-performed baseline (i.e. RTD), the performance gains achieved by SRTD scheme on SPC, MCC, and Kappa are 10.5%, 7.1% and 2.0%, respectively. Figure 4 also shows that both SRTD and the original RTD significantly outperform all other baselines. This is due to the fact that both algorithms are designed to be robust against misinformation spread. In particular, both schemes are able to identify a vast majority of the false rumors in the dataset while other baselines misclassify many popular rumors as truthful information. It is also observed that SRTD outperforms the original RTD scheme, especially in terms of specificity. This is because the fine-grained credibility scores of reports introduced by SRTD allow identification of more false claims than the RTD scheme when data is sparse. This enhancement of the SRTD scheme results in a higher true negative rate and a lower false positive rate.

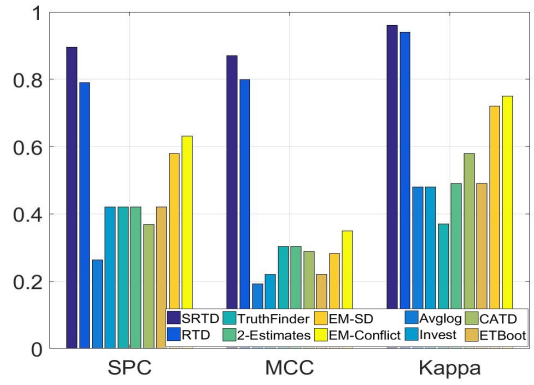


Figure 4. Dallas Shooting Data Trace

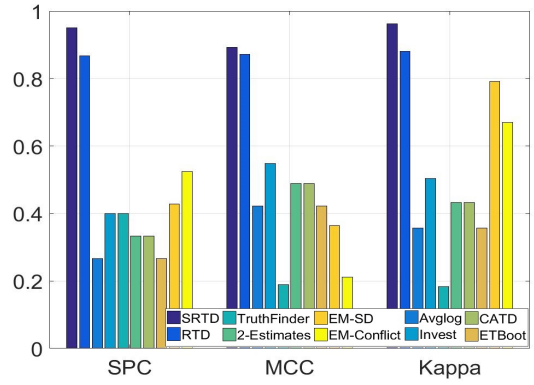


Figure 5. Charlie Hebdo Attack Data Trace

The evaluation results of the Charlie Hebdo Attack data trace are shown in Figure 5. We observe that SRTD continues to outperform all baselines on all three evaluation metrics. In particular, the performance gain achieved by SRTD compared to the best-performing baseline (i.e. RTD) on SPC, MCC, and Kappa is 9.3%, 2.1%, and 8.1%, respectively. The MCC scores for SRTD and RTD are similar because both schemes can identify most false rumors from this data trace.

The results of the Boston Bombing data trace are shown in Figure 6. We observe that the SRTD scheme performs the best among all compared schemes. Compared to RTD (the best-performing baseline), the performance gains achieved by SRTD scheme on SPC, MCC, and Kappa are 7.5%, 3.0% and 7.4% respectively. We also observe that TruthFinder achieves a similar performance to the SRTD scheme on specificity on this dataset. This is because the TruthFinder is prone to output more negative outputs, which lead to a smaller false positive and a large false negative result.

6.2.5 Evaluation Results - Scalability and Efficiency

Next, we evaluate the efficiency of the SRTD scheme. We run SRTD on the Notre Dame HTCondor cluster with a ten worker maximum. We run all other baselines on a single node with 4 processors and 8G of RAM. We use such a small number of workers (i.e., 10 as compared to a total of over 14,700 cores available in the HTCondor system) in order to favor other centralized baselines that are not designed to run in a cluster. This demonstrates the performance gain of our scheme with a very limited amount of resources (i.e., our scheme only needs 10 workers to outperform other baselines). The execution time of all compared schemes on

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2824812, IEEE Transactions on Big Data

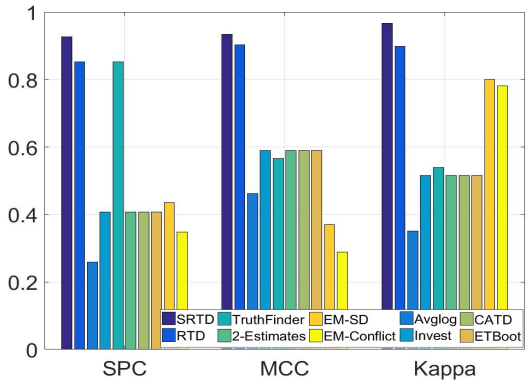


Figure 6. Boston Bombing Data Trace

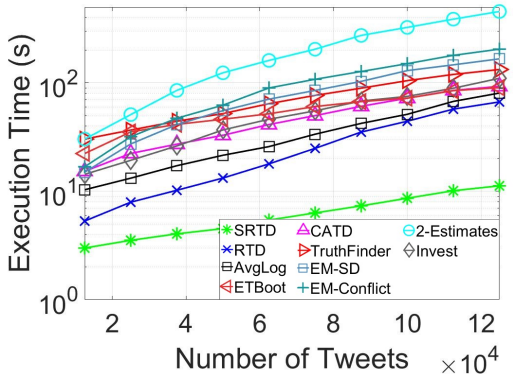


Figure 7. Dallas Shooting Data Trace

Table 3
Running Time (Seconds)

Method	Dallas	Charlie Hebdo	Boston
SRTD	**4.762**	**4.169**	**7.845**
RTD	10.863	11.023	21.376
CATD	32.763	34.846	40.434
TruthFinder	57.519	40.213	82.957
2-Estimates	143.234	152.707	311.941
Invest	76.971	83.827	127.369
AverageLog	21.133	17.922	35.697
ETBOOT	46.223	81.170	54.412
EM-SD	55.772	52.793	62.457
EM-Conflict	62.597	56.214	75.923

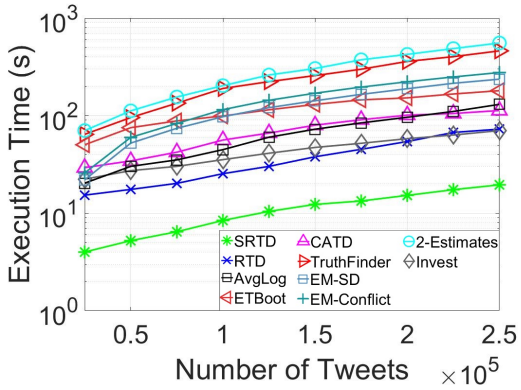


Figure 8. Charlie Hebdo Attack Data Trace

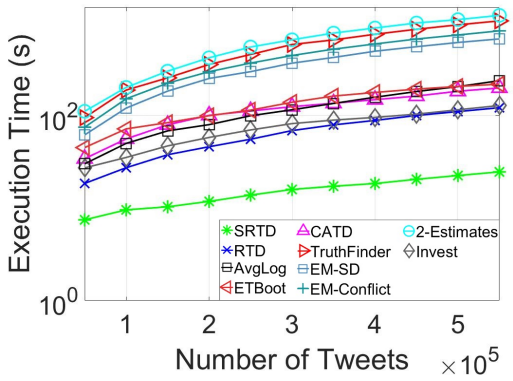


Figure 9. Boston Bombing Data Trace

the three data traces are shown in Table 3. The results show that our SRTD scheme outperforms all other baselines by having a shorter execution time to finish the truth discovery task.

We further evaluate the scalability of the SRTD scheme by extending each of the data traces by adding "synthetic" tweets. In particular, we use the entire original data trace that contains many more tweets than our evaluation set (e.g., tweets contribute to unconfirmed or irrelevant claims, or tweets are in other languages rather English). The size of each data trace for scalability analysis is listed in Table 2. The results are shown in Figure 7-9. We observe that the SRTD scheme is the fastest of all compared schemes as the data size increases. We also observe that the performance gain achieved by SRTD becomes more significant when data size becomes larger. These results demonstrate the scalability of our scheme on large data traces in social media sensing applications. We envision that the performance gain of SRTD over other baselines would be much larger if we ran it with more nodes.

6.2.6 *Evaluation Results - Convergence*

Finally, we perform the convergence analysis of the SRTD scheme. The results are shown in Figure 10. The y-axis is the difference of average claim truthfulness scores between two consecutive iterations. We observe that the SRTD scheme converges after only a couple of iterations on all data traces.

7 Discussion and Future Work

This section discusses some limitations we have identified in the current SRTD scheme as well as the future work that we plan to carry out to address these limitations. First, the SRTD scheme relies on a set of heuristically defined scoring functions to solve the truth discovery problem in social media sensing applications. In particular, the SRTD scheme explores the semantic information of the reports to address the data sparsity problem. However, it would also be interesting to explore the possibility of developing robust and principled truth discovery models that can address the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2824812, IEEE Transactions on Big Data

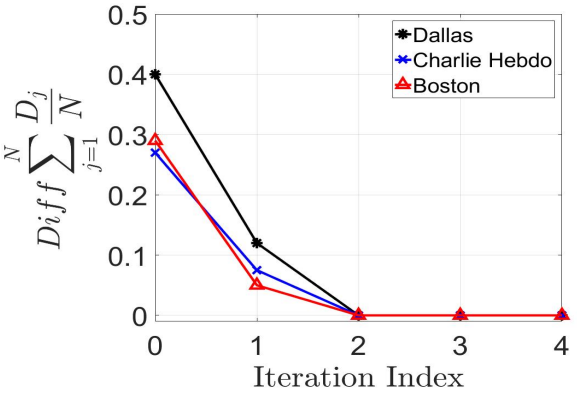


Figure 10. Convergence of SRTD

data sparsity problem with rigorous objective functions and optimized solutions [11]. Specifically, we plan to explore principled statistical models that can explicitly handle the sparse data. For example, a set of estimation theoretical models can be extended to address the data sparsity issue by using a sparse maximum likelihood estimation framework [1]. Alternatively, data fusion techniques can be applied to incorporate external data sources (e.g., traditional news media) to augment the sparse data obtained from social media [46]. Finally, the explicit or implicit dependencies between claims can also be explored under a principled analytical framework to mitigate the data sparsity issue in principled truth discovery solutions [32]. The authors are actively working in this direction.

Second, the proposed scheme does not consider the unconfirmed claims that do not have ground truth or cannot be independently verified by a trustworthy source external to Twitter. However, unconfirmed claims are quite common in real-world social media sensing applications. Compared to the claims whose ground truth can be verified, many unconfirmed tweets simply express personal feelings or "shoutouts". In the future, we plan to address this limitation by identifying and filtering out unconfirmed claims using current sentiment analysis methods on social media [2], [21]. Alternatively, we can also extend our SRTD scheme by generalizing the categories of claims to include the unconfirmed ones. The challenge lies in defining credibility scores of reports related to the unconfirmed claims and integrating them into the new SRTD scheme.

Third, SRTD does not consider the dynamic truth problem where the truth of a claim changes over time (e.g., escape path of a suspect). There are two critical tasks in addressing the dynamic truth challenge. The first is to capture the transition of truth promptly, especially in a time critical system, where we need to detect the updates of true information frequently (e.g. the update of a terrorist attack, the current stock price). The second one is to be robust against noisy data that may lead to an incorrect detection of the truth transition. Such a task can be challenging especially in a social media sensing scenario where sources can easily spread rumors. We plan to address this problem by integrating a Hidden Markov Model (HMM) [24] with our SRTD scheme to capture such dynamics based on the observed crowd opinion as well as the previous states of truth. In particular, we can treat the truthfulness of claims as the hidden states and contribution scores of tweets as observations. We can then leverage the HMM to explicitly model the state transition probability of claim truthfulness as well as the emission probability of crowd opinions. The Viterbi algorithm [30] can be applied to decode the dynamic truth of claims from the HMM model.

Fourth, a common limitation for our scheme and other truth discovery techniques is that false claims can spread from one domain to another domain without changing any information, making it hard to identify truthful information in real time. For example, during the Boston Bombing event, CNN claimed that a bomber was arrested two days after the event. This original message was retweeted more than 3,000 times until, half an hour later, it was debunked by the Boston police department claiming that no arrest had been made. Without any debates before the debunking, our scheme may fail to detect that such a rumor is false. However, such scenarios normally appear in the early stage of an event, and eventually, debates and questions about the rumor will pop out. In fact, in our datasets, all of the rumors have conflicting opinions (debates) within the duration of the events, which provides us with a solid basis to detect false claims. Also, our scheme degrades the importance of non-original claims (e.g., repeated or simply forwarded ones), which provides robustness against misinformation spread.

Fifth, the current SRTD scheme assumes independence between claims. There may be cases, however, when one claim could be related to other claims (e.g., weather conditions at city B may be related to weather conditions at city A when A and B are close in distance). Claim dependency is sometimes implicit and requires extra domain knowledge. For example, the claims "OSU student shot and killed near campus" and "a car w\2ppl rammed Watts Hall. 1 w\knife 1 w\gun." are actually correlated given the fact that Watts Hall is a building inside the OSU campus. However, without such extra domain knowledge, it would be difficult or even impossible to identify such dependencies between claims. Incorporating such dependencies into the SRTD framework can be an interesting topic for future research. In the future, we plan to use a lexical database such as WordNet[4] to explicitly model the relationships between words on similar concepts. We can also model the physical dependencies between claims based on their geotagged locations using location-based services such as Google Maps[5].

Finally, we also plan to further enhance the robustness of the SRTD scheme against collusion attacks where a group of users may intentionally generate and propagate misinformation to mislead the crowd (e.g., to influence a political election or to spam users). Unfortunately, this problem has not been well addressed in current truth discovery solutions. We plan to address this problem by explicitly incorporating the dependencies among potentially colluded sources into the SRTD scheme. Such dependencies can be estimated based on similarities between tweets, frequency and timing of tweets, and the following/followee relationship between Twitter users. Social honey-pot techniques [15]

4. https://wordnet.princeton.edu/
5. https://developers.google.com/maps/

can also be used to actively analyze the properties of the attacker's profiles by luring them to debut on some carefully crafted claims. Such profile information could be further incorporated into the source reliability computation of our SRTD scheme.

8 Conclusion

In this paper, we proposed a Scalable Robust Truth Discovery (SRTD) framework to address the data veracity challenge in big data social media sensing applications. In our solution, we explicitly considered the source reliability, report credibility, and a source's historical behaviors to effectively address the misinformation spread and data sparsity challenges in the truth discovery problem. We also designed and implemented a distributed framework using Work Queue and the HTCondor system to address the scalability challenge of the problem. We evaluated the SRTD scheme using three real-world data traces collected from Twitter. The empirical results showed our solution achieved significant performance gains on both truth discovery accuracy and computational efficiency compared to other state-of-the-art baselines. The results of this paper are important because they provide a scalable and robust approach to solve the truth discovery problem in big data social media sensing applications where data is noisy, unvetted, and sparse.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. CBET-1637251, CNS-1566465 and IIS-1447795, Google Faculty Research Award 2017, and Army Research Office under Grant W911NF-17-1-0409. 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] O. Banerjee, L. E. Ghaoui, and A. dAspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516, 2008.

[2] S. Bhuta and U. Doshi. A review of techniques for sentiment analysis of twitter data. In *Proc. Int Issues and Challenges in Intelligent Computing Techniques (ICICT) Conf*, pages 583–591, Feb. 2014.

[3] J. Bian, Y. Yang, H. Zhang, and T.-S. Chua. Multimedia summarization for social events in microblog stream. *IEEE Transactions on multimedia*, 17(2):216–228, 2015.

[4] P. Bui, D. Rajan, B. Abdul-Wahid, J. Izaguirre, and D. Thain. Work queue+ python: A framework for scalable scientific ensemble applications. In *Workshop on python for high performance and scientific computing at sc11*, 2011.

[5] P.-T. Chen, F. Chen, and Z. Qian. Road traffic congestion monitoring in social media with hinge-loss markov random fields. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 80–89. IEEE, 2014.

[6] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. In *Proceedings of the VLDB Endowment*, pages 550–561, 2009.

[7] X. X. et al. Towards confidence in the truth: A bootstrapping based truth discovery approach. In *Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 2016.

[8] R. Farkas, V. Vincze, G.Mora, J. Csirik, and G.Szarvas. The conll-2010 shared task: Learning to detect hedges and their scope in natural language text. In *In Proceedings of the Fourteenth Conference on Computational Natural Language Learning.*, 2010.

[9] R. Feldman and M. Taqqu. A practical guide to heavy tails: statistical techniques and applications. *Springer Science & Business Media*, 1998.

[10] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *In Proc. of the ACM International Conference on Web Search and Data Mining (WSDM'10)*, pages 131–140, 2010.

[11] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014.

[12] 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.

[13] A. Karandikar. *Clustering short status messages: A topic model based approach*. PhD thesis, University of Maryland, Baltimore County, 2010.

[14] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The web as a graph: Measurements, models, and methods. In *International Computing and Combinatorics Conference*, pages 1–17. Springer, 1999.

[15] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 435–442. ACM, 2010.

[16] M. J. Litzkow, M. Livny, and M. W. Mutka. Condor-a hunter of idle workstations. In *Distributed Computing Systems, 1988., 8th International Conference on*, pages 104–111. IEEE, 1988.

[17] J. Marshall and D. Wang. Mood-sensitive truth discovery for reliable recommendation systems in social sensing. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 167–174. ACM, 2016.

[18] E. Mustafaraj, S. Finn, C. Whitlock, and P. T. Metaxas. Vocal minority versus silent majority: Discovering the opionions of the long tail. In *Proc. IEEE Third Int Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conf. Social Computing (SocialCom) Conf*, pages 103–110, Oct. 2011.

[19] nielson. Super bowl 50: Nielsen twitter tv ratings post-game report.

[20] R. W. Ouyang, L. M. Kaplan, A. Toniolo, M. Srivastava, and T. Norman. Parallel and streaming truth discovery in large-scale quantitative crowdsourcing.

[21] R. Pandarachalil, S. Sendhilkumar, and G. S. Mahalakshmi. Twitter sentiment analysis for large-scale data: An unsupervised approach. *Cognitive Computation*, 7(2):254–262, Nov. 2014.

[22] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). *Association for Computational Linguistics*, pages 877–885, 2010.

[23] J. Qadir, A. Ali, A. Zwitter, A. Sathiaseelan, J. Crowcroft, et al. Crisis analytics: Big data driven crisis response. *arXiv preprint arXiv:1602.07813*, 2016.

[24] L. Rabiner and B. Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

[25] A. Rangrej, S. Kulkarni, and A. V. Tendulkar. Comparative study of clustering techniques for short text documents. In *Proceedings of the 20th international conference companion on World wide web*, pages 111–112. ACM, 2011.

[26] K. Starbird, J. Maddock, M. Orand, P. Achterman, and R. M. Mason. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. In *iConference 2014 Proceedings*, 2014.

[27] D. Thain and C. Moretti. Abstractions for cloud computing with condor. 2010.

[28] D. Thain, T. Tannenbaum, and M. Livny. Condor and the grid. *Grid computing: Making the global infrastructure a reality*, pages 299–335, 2003.

[29] M. Y. S. Uddin, M. T. A. Amin, H. Le, T. Abdelzaher, B. Szymanski, and T. Nguyen. On diversifying source selection in social sensing. In *Proc. Ninth Int Networked Sensing Systems (INSS) Conf*, pages 1–8, June 2012.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2824812, IEEE Transactions on Big Data

[30] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, Apr. 1967.

[31] D. Wang, T. Abdelzaher, and L. Kaplan. *Social sensing: building reliable systems on unreliable data*. Morgan Kaufmann, 2015.

[32] D. Wang, T. Abdelzaher, L. Kaplan, R. Ganti, S. Hu, and H. Liu. Exploitation of physical constraints for reliable social sensing. In *Real-Time Systems Symposium (RTSS), 2013 IEEE 34th*, pages 212–223. IEEE, 2013.

[33] D. Wang, T. Abdelzaher, L. Kaplan, R. Ganti, S. Hu, and H. Liu. Reliable social sensing with physical constraints: analytic bounds and performance evaluation. *Real-Time Systems*, 51(6):724–762, 2015.

[34] D. Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. C. Aggarwal, R. Ganti, X. Wang, P. Mohapatra, B. Szymanski, and H. Le. Using humans as sensors: An estimation-theoretic perspective. In *Proc. 13th Int Information Processing in Sensor Networks Symp. IPSN-14*, pages 35–46, Apr. 2014.

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

[36] D. Wang, L. Kaplan, and T. F. Abdelzaher. Maximum likelihood analysis of conflicting observations in social sensing. *ACM Transactions on Sensor Networks (ToSN)*, 10(2):30, 2014.

[37] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher. On truth discovery in social sensing: A maximum likelihood estimation approach. In *Proc. ACM/IEEE 11th Int Information Processing in Sensor Networks (IPSN) Conf*, pages 233–244, Apr. 2012.

[38] D. Wang, B. K. Szymanski, T. Abdelzaher, H. Ji, and L. Kaplan. The age of social sensing. *arXiv preprint arXiv:1801.09116*, 2018.

[39] Z. Xu, Y. Liu, N. Yen, L. Mei, X. Luo, X. Wei, and C. Hu. Crowdsourcing based description of urban emergency events using social media big data. *IEEE Transactions on Cloud Computing*, 2016.

[40] W. Xue, J. Shi, and B. Yang. X-rime: cloud-based large scale social network analysis. In *Services Computing (SCC), 2010 IEEE International Conference on*, pages 506–513. IEEE, 2010.

[41] S. R. Yerva, H. Jeung, and K. Aberer. Cloud based social and sensor data fusion. In *Information Fusion (FUSION), 2012 15th International Conference on*, pages 2494–2501. IEEE, 2012.

[42] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808, June 2008.

[43] X. Yin and W. Tan. Semi-supervised truth discovery. In *Proceedings of the 20th international conference on World wide web*, number 217-226. ACM, 2011.

[44] D. Zhang, H. Rungang, and D. Wang. On robust truth discovery in sparse social media sensing. In *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016.

[45] D. Y. Zhang, Y. Ma, Y. Zhang, S. Lin, X. S. Hu, and D. Wang. A real-time and non-cooperative task allocation framework for social sensing applications in edge computing systems. In *Real-Time and Embedded Technology and Applications Symposium (RTAS), 2018 IEEE*. IEEE, 2018. accepted.

[46] D. Y. Zhang, D. Wang, and Y. Zhang. Constraint-aware dynamic truth discovery in big data social media sensing. In *Big Data (Big Data), 2017 IEEE International Conference on*, pages 57–66. IEEE, 2017.

[47] D. Y. Zhang, D. Wang, H. Zheng, X. Mu, Q. Li, and Y. Zhang. Large-scale point-of-interest category prediction using natural language processing models. In *Big Data (Big Data), 2017 IEEE International Conference on*, pages 1027–1032. IEEE, 2017.

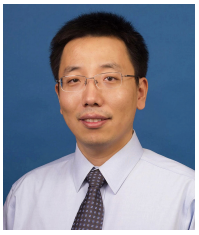
[48] D. Y. Zhang, C. Zheng, D. Wang, D. Thain, X. Mu, G. Madey, and C. Huang. Towards scalable and dynamic social sensing using a distributed computing framework. In *Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on*, pages 966–976. IEEE, 2017.

[49] J. Zhang and D. Wang. Duplicate report detection in urban crowdsensing applications for smart city. In *Smart City/SocialCom/SustainCom (SmartCity), 2015 IEEE International Conference on*, pages 101–107. IEEE, 2015.

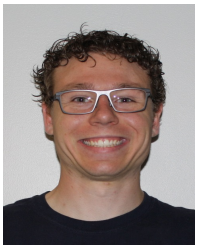
[50] Y. Zhang. A cross-site study of user behavior and privacy perception in social networks. Master's thesis, Purdue University, 2014.

[51] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. In *Proceedings of the VLDB Endowment*, volume 5, pages 550–561, 2012.

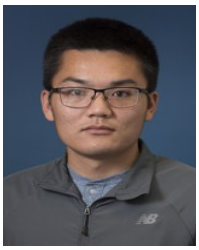
Daniel (Yue) Zhang is a PhD student in the Department of Computer Science and Engineering at the University of Notre Dame. He received his M.S. degree from Purdue University, West Lafayette, Indiana, USA, in 2012 and a B.S. degree from Shanghai Jiao Tong University, Shanghai, China, in 2008. His research interests include human-centric computing, social sensing based edge computing ecosystem, truth analysis on social media, and Cyber-Physical Systems. He is a student member of IEEE.



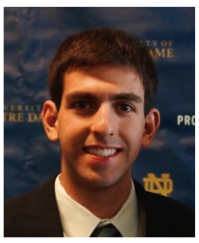
Dong Wang received his Ph.D. in Computer Science from University of Illinois at Urbana Champaign (UIUC) in 2012, an M.S. degree from Peking University in 2007 and a B.Eng. from the University of Electronic Science and Technology of China in 2004, respectively. He is now an assistant professor in the Department of Computer Science and Engineering at the University of Notre Dame. Dr. Wang's research interests lie in the area of reliable social sensing, cyber-physical computing, real-time and embedded systems, and crowdsourcing applications. He received the Google Faculty Research Award in 2018, Army Research Office Young Investigator Program (YIP) Award in 2017, Wing-Kai Cheng Fellowship from the University of Illinois in 2012 and the Best Paper Award of IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS) in 2010. He is a member of IEEE and ACM.



Nathan Vance is a PhD student in the Department of Computer Science and Engineering at the University of Notre Dame. He received his B.S. degree in Computer Science at Hope College in Holland, Michigan, in 2017. His research focuses on edge computing, natural language processing, and security.



Yang Zhang is a PhD student in the Department of Computer Science and Engineering at the University of Notre Dame. His research interests include transfer learning, case-based reasoning, optimized task allocation in social sensing, and multi-view data fusion.



Steven Mike is an undergraduate student at the University of Notre Dame. He works in the Social Sensing lab as an undergraduate research assistant. His research interests mainly focus on data fusion in online social media and emerging applications in mobile social sensing.