Towards Reliable Missing Truth Discovery in Online Social Media Sensing Applications

Daniel (Yue) Zhang, Jose Badilla, Yang Zhang, Dong Wang
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN, USA
yzhang40@nd.edu, jbadilla@nd.edu, yzhang42@nd.edu, dwang5@nd.edu

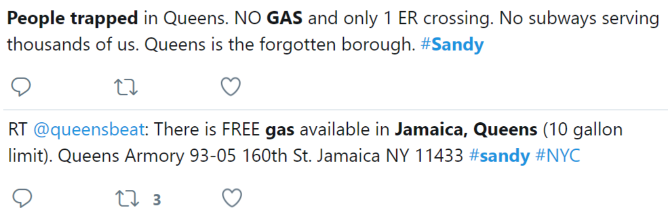
***Abstract*—Social media sensing has emerged as a new application paradigm to collect observations from online social media users about the physical environment. A fundamental problem in social media sensing applications lies in estimating the evolving truth of the measured variables and the reliability of data sources without knowing either of them *a priori*. This problem is referred to as *dynamic truth discovery*. Two major limitations exist in current truth discovery solutions: i) existing solutions cannot effectively address the *missing truth* problem where the measured variables do not have any reported measurements from the data sources; ii) the latent correlations among the measured variables were not fully captured and utilized in current solutions. In this paper, we proposed a Reliable Missing Truth Finder (RMTF) to address the above limitations in social media sensing applications. In particular, we develop a novel data-driven technique to identify the lagged and latent correlations among measured variables, and incorporate such correlation information into a holistic spatiotemporal inference model to infer the missing truth. We evaluated the RMTF using the real-world Twitter data feeds. The results show that the RMTF scheme significantly outperforms the state-of-the-art truth discovery solutions by correctly inferring the missing truth of the measured variables.**

***Index Terms*—Missing Truth Discovery, Social Media Sensing, Spatiotemporal Inference**

I. Introduction

Social media sensing has become a new application paradigm to collect the measurements (often called claims) about the physical world from the observations reported by users on social media [1], [2], [3], [4]. Examples of social sensing applications include real-time traffic monitoring using mobile apps [5], obtaining real-time situation awareness during disaster events via online social media [6], and object tracking using portable video devices [7]. A critical challenge in social media sensing is to accurately discover the evolving truth of the measured variables from the massive noisy measurements contributed by unreliable human sensors [8], [9]. We refer to this problem as *dynamic truth discovery*. For example, in the aftermath of a natural disaster (e.g., earthquake, hurricane), human sensors (e.g., Twitter users) post reports on social media about their observations of the disaster (e.g., impacted areas, the number of casualties, locations of the available resource) in real-time. It is crucial to reliably identify the truthful information from social media sensing applications for effective decision makings. Recent efforts have been made to solve the dynamic truth discovery problem [9], [10], [2]. Examples of these solutions include Markov models [9], [11], [12], Bayesian networks [13], [14] and maximum likelihood estimation (MLE) methods [10], [15], [16]. However, two critical challenges have not been fully addressed: *missing truth* and *lagged latent correlation*.

Missing Truth: the social media sensing data is observed to be sparse due to the spontaneous nature of human sensors [17]. For example, the human sensors (e.g., Twitter users) may lack the motivation and incentives to continuously contribute data to the application [12]. Alternatively, they may also choose to ignore topics or events they are not interested in and only contribute data to the topics or events that match their interests [8]. For example, consider a "Gas Finder" application where social media users collectively report potentially conflicting claims about the gas availability at different stations (measured variables) in the aftermath of a hurricane (Figure 1). It is unlikely that the human sensors would constantly report the availability of all gas stations during the entire period of the hurricane. In fact, we found that an average of 45% of gas stations did not have reports from any user on a daily basis in the dataset we collected from Twitter. Such missing truth problem can lead to the truth discovery results that erroneously guide the drivers who are eagerly searching for gas to a sold-out station. Existing truth discovery solutions often assume that a measured variable is constantly reported by a dense set of data sources and fail to address the missing truth challenge [18].


People trapped in Queens. NO **GAS** and only 1 ER crossing. No subways serving thousands of us. Queens is the forgotten borough. **#Sandy**

RT @queensbeat: There is FREE **gas** available in **Jamaica, Queens** (10 gallon limit). Queens Armory 93-05 160th St. Jamaica NY 11433 **#sandy** #NYC

Figure 1: Conflicting Claims from Twitter Users Posted on Nov. 3rd 2012 Regarding the Gas Availability During Hurricane Sandy

Lagged latent correlation: correlations among measured variables can be explored to address the missing truth challenge [19]. For example, the availability of gas in a station may

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00 © 2018 IEEE

be highly correlated with other stations in a close proximity or share the same supply chain. However, identifying the correlations among measured variables is a non-trivial task due to several reasons. First, the correlations among measured variables can be latent and may not be directly observable from the collected sensing measurements [20]. Second, the lagged correlation is also quite common in social media sensing applications. In the gas station example, a gas station that is running out of gas will prompt people around it to seek nearby gas stations, causing those be sold-out as well. This results in a delayed (lagged) indirect correlation. Current truth discovery solutions either completely ignore the correlations between measured variables [18], [2] or assume such correlations to be known as prior knowledge [19]. Therefore, there is a lack of a principled analytical framework that can explore the lagged and latent correlation of measured variables and incorporate such correlation in solving the dynamic truth discovery problem.

To address the above challenges, this paper develops a new Reliable Missing Truth Finder (RMTF) scheme to solve the dynamic truth discovery problem with missing truth and latent lagged correlation between measured variables. In particular, we develop a dynamic correlation inference module to identify the latent correlation among measured variables inspired by the dynamic mixture topic modeling techniques from text mining. We then develop a principled framework to reliably infer the missing truth of the measured variables by leveraging the inferred correlation between them. The evaluation results on a real-world Twitter dataset demonstrates that the RMTF significantly outperforms existing truth discovery solutions by correctly inferring the missing truth of the measured variables.

II. Related Work

A. Social Media Sensing and Truth Discovery

Social media sensing has received a significant amount of attention due to the increasing popularity of smart devices (e.g., smartphones, tablets), the advent of online data sharing platforms (e.g., online social media), and the proliferation of Internet connectivity (e.g., WiFi, 5G) [21]. Examples of social media sensing applications include intelligent transportation systems [22], urban sensing [23], personalized recommendation system [12], copyright infringement detection [15], [24] and disaster and emergency response [2]. A few important challenges exist in social media sensing applications [25]. Examples include data reliability [8], incentive mechanism [3], task allocation [26], sensor profiling [27], [28], heterogeneous data fusion [29], and security and privacy preservation [30].

A critical problem in social media sensing is the *Truth Discovery* problem, of which the goal is to identify the truthful information among unreliable social media data [2], [10], [31], [32]. The truth discovery problem has received an increasing amount of attention given the spread of fake news, spams, and misinformation on social media. A comprehensive survey of truth discovery solutions in social media sensing applications is given in [25]. Currently, a trending topic in the truth discovery research is the "dynamic truth discovery", where the ground truth of the measured variables can change over time [9]. For example, Zhang *et al.* developed a constraint-aware truth discovery model to incorporate physical constraints into the detection of the evolving truth [11]. Zhao *et al.* proposed a single-pass truth discovery method to explicitly consider memory constraints and computation efficiency in performing truth discovery on streaming data [33]. Wang *et al.* proposed a dynamic truth discovery scheme that could jointly estimate source reliability and the correctness of claims using a recursive maximum-likelihood estimation approach [10]. However, two major limitations exist in these approaches. First, they did not address the problem of missing truth where some measured variables have no reported claims at a certain point of time. Second, the correlations among measured variables were either completely ignored or assumed to be known *a priori* [17], [34]. In this paper, we develop a RMTF scheme to address these limitations.

B. Spatiotemporal Inference

Spatiotempoal inference models are also related to this research. To capture both temporal and spatial dependencies of measured variables, a set of models were proposed to extend the traditional autoregressive model to incorporate both spatial and temporal correlations. These models include STARIMA model proposed by Kamarianakis *et al.* to predict urban traffic flow [35], STAR model proposed by Pace *et al.* to predict real estate price [36] and STARMAX model proposed by Pace *et al.* to infer fishery data [37]. A set of more recent machine learning based approaches were also developed to perform spatiotemporal inference [38]. While these solutions motivate our work, they have several limitations: i) they assume static spatial correlation between measured variables, which does not hold in many social media sensing applications [9]; ii) these approaches often ignore the unique data reliability issue in social media sensing where human sensors can post conflicting claims. In contrast, we develop a RMTF framework that explicitly considers ii) the underlying spatio-temporal dynamics of the measured variables, and ii) the conflicting information from the social media users.

C. Infer Unobserved Sensor Data

Previous studies have made great progress to infer the missing sensor data to obtain desirable spatiotemporal coverage in crowdsourcing applications [39], [23], [40]. For example, Wang *et al.* proposed a compressive sensing based approach to deduce the missing data of unmeasured areas [39]. Hsieh *et al.* developed a semi-supervised inference algorithm to infer the air quality index of a city by exploring the existing monitoring data and external information such as points of interest, road structure, and human mobility [23]. Umer *et al.* proposed a localized and distributed spatial interpolation scheme to infer the data from areas that do not have sensor coverage [40]. Although the above works share the similar goal as our work of inferring the missing data due to the sparse data coverage, they did not explicitly consider the unique *data reliability* issue where data from social media may contain conflicting information. In contrast, this data reliability problem is addressed in our proposed framework.

III. Problem Definition

In this section, we present the problem of dynamic truth discovery with missing truth in social media sensing applications. We first define a few key terms.

DEFINITION 1. Measured Variable: A measured variable is a variable of interest in a social media sensing application.

DEFINITION 2. Source: A source is a social media user who reports his/her observation of the measured variables.

DEFINITION 3. Claim: A claim is the reported observation from a source on a measured variable.

DEFINITION 4. Sensing Cycle: A sensing cycle is a period of time during which reported claims are collected and analyzed. All claims collected in the same sensing cycle are considered to be "concurrent."

Figure 2 shows the "Gas Finder" application introduced in Section I. In this example, a measured variable is the gas availability of a gas station, a source is a Twitter user and a claim is a tweet that discusses the gas availability of a station. The sensing cycle is the time period (e.g., a day) defined by the application to update the availability of the gas stations. Note that some gas stations may have no reported claims at a certain sensing cycle. The true values of these measured variables are considered as *missing truth* accordingly. To formally define our problem, we consider a social media sensing application with M measured variables and T sensing cycles. A set of N^t sources collectively report claims of the measured variables at the t^{th} sensing cycle. We use x_i^t to denote the true value of the i^{th} measured variable and c_i^t to denote all the claims contributed to the i^{th} measured variable at the t^{th} sensing cycle. We further define $c(t) = \{c_1^t, c_2^t, ..., c_M^t\}$ as the claims of all measured variables at the t^{th} sensing cycle.

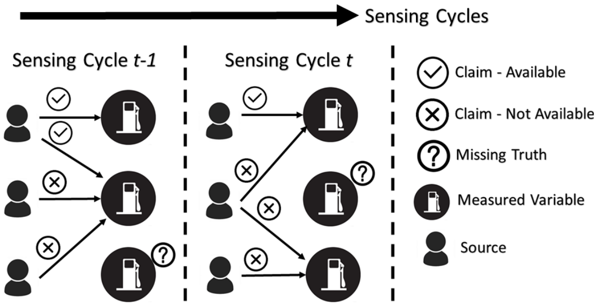


Figure 2: Dynamic Truth Discovery with Missing Truth

The goal of our problem is to infer the true values of the measured variables in the current sensing cycle (i.e., t) given the claims contributed by all sources. Formally, we compute:

$$\arg\max_{\tilde{x}(t)} Pr(\tilde{x}(t) = x(t)|c(1), c(2), ..., c(t)), \forall\ 1 \le t \le T \tag{1}$$

where $x(t) = \{x_1^t, x_2^t, ...x_M^t\}$ and $\tilde{x}(t) = \{\tilde{x}_1^t, \tilde{x}_2^t, ...\tilde{x}_M^t\}$ denote the ground truth and estimated values of all measured variables at the t^{th} sensing cycle, respectively.

IV. Solution

In this section, we present the RMTF framework to address the problem formulated in the previous section. In the following subsections, we first discuss how we capture the correlations between measured variables. Then we develop a new truth discovery model that can take the derived correlations to infer the true values of measured variables.

A. Dynamic Correlation Inference Module

We first present a novel Dynamic Correlation Inference (DCI) module to infer the correlations among measured variables using dynamic mixture topic modeling techniques. In our approach, we capture the "lagged and latent correlations" among measured variables and dynamically update such correlations as new claims arrive. We do not assume any prior knowledge or domain-specific information about the underlying features that govern the evolving truth of measured variables; we thereby keep our approach generic.

1) Infer Lagged and Latent Variable Correlation: The DCI module infers the correlations based on the similarities of the *latent features* of the measured variables. In the gas availability example, the availability of different gas stations may be correlated due to features such as location, supply chain, company brand, scale of the gas station, etc. These features cannot be enumerated, and are therefore treated as latent features in our model. These latent features can be captured by the dynamic mixture topic modeling technique [20] where a sequence of observed values (referred to as a "truth sequence") of a measured variable are assumed to be generated by a set of latent topics. Different from the static topic models, our dynamic mixture topic model assumes that the topic distribution of the measured variables can evolve over time. We assume that the generation process of the observed true values is as follows:

1) Draw binomial distributions θ^1 and Φ_z with Dirichlet parameters α, β respectively at the 1^{st} sensing cycle.
2) For each cycle t, draw θ^t from Gaussian distribution with expectation θ^{t-1}. Then draw a latent feature z from a binomial distribution with parameter θ^t.
3) Draw a new true value from a binomial distribution with parameter Φ_z. Sample a multinomial distribution θ^1 from a Dirichlet distribution with parameter α. Draw a multinomial distribution Φ_z for each latent feature z from a Dirichlet distribution with parameter β.

Some defined notations are summarized in Table I.

The likelihood of generating the observed values of measured variables can thus be derived as:

$$
\begin{aligned}
\mathcal{L} &= Pr(S^1, S^2, ...S^T|\alpha, \beta) \\
&= \int\int \prod_{z=1}^{K} Pr(\Phi_z|\beta) \times Pr(\theta^1|\alpha) \prod_{t=2}^{T} Pr(\theta^t|\theta^{t-1}) \\
&\prod_{t=1}^{T}\prod_{n=1}^{L}\sum_{z^t(n)=1}^{K} (Pr(z^t(n)|\theta^t)
\end{aligned}
\tag{2}
$$

The evolving latent topic distribution can be derived using the LDA Gibbs Sampling technique described in [20].

Table I: Notations

Symbol	Description
d	a stream of a sensor's measurements
z	a latent feature, represented as an integer ∈ [1, K]
K	total number of latent features
S^t	a truth sequence at cycle t
L	length of the truth sequence
$x^t(n)$	the n^{th} true value in S^t
$z^t(n)$	the latent feature for the n^{th} true value in S^t
θ^t	mixture distribution of latent features at cycle t
Φ_z	mixture distribution over true values for a latent feature
α, β	hyperparameters (Dirichlet) for θ and Φ respectively

Given the latent topic distributions of two measured variables, we can compute the correlation between them as:

$$
Cor_q^t(i,j) = 1 - \delta(D_{KL}^{sym}(Pr_i^t(Z), Pr_j^{t-q}(Z)) \tag{3}
$$

$$
\begin{aligned}
D_{KL}(P_i^t(Z), Pr_j^{t-q}(Z)) = &\sum_{z\in Z} Pr_i^t(z) log \frac{Pr_j^{t-q}(z)}{Pr_i^t(z)} \\
&+ \sum_{z\in Z} Pr_j^{t-q}(z) log \frac{Pr_i^t(z)}{Pr_j^{t-q}(z)}
\end{aligned}
\tag{4}
$$

where $Pr_i^t(Z)$ and $Pr_j^t(Z)$ denote the latent feature distributions of the i^{th} and j^{th} measured variables at the t^{th} sensing cycle. $Cor_q^t(i,j)$ is the correlation between the two measured variables with lag $q, 1 \leq q \leq Q$. $D_{KL}^{sym}(Pr_i^t(Z), Pr_j^{t-q}(Z))$ is the symmetric KL-Divergence [41] between the two measured variables (with time lag q). δ is a normalization process to map the divergence to a [0,1] scale. The intuition is that two measured variables that share very similar latent feature distributions are more likely to have a stronger spatial correlation.

2) Leveraging Temporal Correlation with ARMA: The measurements of each cell form a time series and the dependencies between consecutive measurements can be helpful in inferring the missing truth. Therefore, in the proposed solution, we also consider the temporal correlations of the measured variables[1]. For example, if a gas station is constantly available in the previous sensing cycles, it is also likely to be available in the current cycle. To capture such temporal correlation, we develop a temporal inference model inspired by the autoregressive–moving-average (ARMA) to estimate the current true value of a measured variable based on the average true values of the previous sensing cycles. Let the order of the ARMA model be P. The estimated true value based on the ARMA model is denoted as: $ARMA_i^t(P) = \sum_{p=1}^{P}(\Lambda_i \tilde{x_i}^{t-p})/P$. Λ_i is a scalar that governs the weights of historical measurements x_i (more recent measurements are assigned higher weights). Note that the dynamic mixture model requires the inference of the missing truth of the measured variables in previous sensing cycles, which is discussed in the next subsection. We assume the correlations among measured variables do not change between two consecutive sensing cycles. Therefore, we can derive the correlations based on the fully estimated true values in previous sensing cycle.

B. Dynamic Truth Discovery with Missing Truth

The key idea of the proposed solution to solve the dynamic truth discovery problem with missing truth is to infer the true value of a measured variable based on i) its historical values, ii) the values of correlated measured variables, and iii) claims from social media users on the measured variable. In our model, we first define a loss function based on the estimated true values and the reported claims of a measured variable as follows.

$$
\begin{aligned}
\mathcal{L}^t = \sum_{i=1}^{N^t} \Bigg\{ &\sum_{c\in c_i^t} w_c \cdot |c - \tilde{x_i}(t)| + \lambda_1 \cdot \sum_{q=1}^{Q}\sum_{j=1}^{N^{t-q}} Cor_q^t(i,j) \cdot \\
&|\tilde{x_j}^{t-q} - \tilde{x_i}^t| + \lambda_2 \cdot |ARMA_i^t(P) - \tilde{x_i}^t| \Bigg\} \\
&s.t. \quad \tilde{x_i}^t \in \{0, 1\}
\end{aligned}
\tag{5}
$$

We refer to the first term $\sum_{c\in c_i^t} w_c \cdot |c - \tilde{x_i}(t)|$ as "claim disagreement" which represents the disagreement between the reported values (claims) and the estimated true values of the measured variables. w_c is the weight assigned to a claim to represent its importance. The weights are derived based on i) whether the claim is independently made (higher weight) or simply copied from others; and ii) whether the claim is assertive (higher weight) or uncertain. Unlike existing truth discovery methods that mainly focus on estimating the reliability of sources [25], our scheme addresses the data sparsity problem in social media where each user may only contribute a limited number of claims [8]. In fact, we found that each user only posted 1.23 claims on average in our collected dataset, which provides insufficient evidence to estimate the reliability of data sources. Therefore, instead of relying on estimating the weights of sources, we focus on estimating the weight of each individual claim. We refer to the second term $\lambda_1 \cdot \sum_{q=1}^{Q}\sum_{j=1}^{N^{t-q}} Cor_q^t(i,j) \cdot |\tilde{x_j}^{t-q} - \tilde{x_i}^t|$ as "spatial disagreement" which represents the disagreement of the estimated true value of a measured variable and the values of its correlated variables. The correlations are normalized so that $\sum_{q=1}^{Q}\sum_{j=1}^{N^{t-q}} Cor_q^t(i,j) = 1$. Similarly, we refer to the third term $\lambda_2 \cdot |ARMA_i^t(P) - \tilde{x_i}^t|$ as "temporal disagreement" which represents the disagreement of the estimated true value of a

[1] We use "spatial" to refer to the correlations between measured variables and "temporal" to represent the between the correct true value and values in previous sensing cycles of the same measured variable.

measured variable and the variable's historical values. λ_1 and λ_2 are hyper parameters that balance the above three terms. The true values of all measured variables can be estimated by minimizing the above "disagreements" (i.e., loss function) using the Binary Integer programming [42] approach.

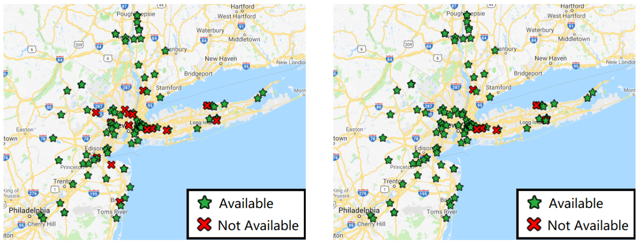
V. Evaluation

In this section, we evaluate RMTF in comparison with the state-of-the-art truth discovery solutions on a real-world dataset collected from Twitter. The results demonstrate that RMTF significantly outperforms all compared baselines.

A. Experimental Setups

1) Dataset and Pre-processing: We evaluate the RMTF based on a typical social media sensing application scenario where social media posts from humans are leveraged to identify the availability of the critical resources during disaster events. We use a Twitter dataset we collected about the gas availability of stations in the states of New Jersey and New York in the aftermath of Hurricane Sandy [2]. This dataset contains a total of 904,362 tweets. To extract information related to gas availability, we filter the tweets using the hashtags. We summarize the hashtags and pre-processed dataset in Table II.

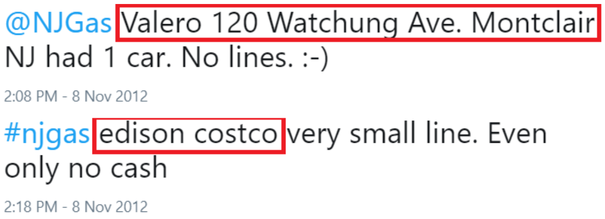
The ground truth of the gas availability at relevant stations is collected from the All Hazards Consortium (AHC) Private Sector Resource Report [3]. This report contains the resource availability information on gas, hotel, pharmacy, food, as well as the GPS locations of these resources in November 2012. For evaluation, we use the data from Nov. 2nd to Nov. 15th (14 days) that overlaps with our Twitter data collection. We observe that the gas availability in the affected area is highly dynamic during the studied period (Figure 3). We set the sensing cycle as 1 day to match the ground truth data.



(a) Gas Availability Nov. 3rd 2012 (b) Gas Availability Nov. 10th 2012

Figure 3: Gas Availability during Hurricane Sandy

For each tweet, we identify the measured variable (i.e., the gas station) by matching the street name and city from the tweet with the locations of stations in the ground truth file (see Figure 4). We discard the tweets with missing location information, as well as those that cannot be mapped to any gas station in the ground truth file. To interpret the "value" of a claim (i.e., gas availability), we classify the tweets using a bag-of-words approach. For example, if the tweet contains a keyword such as "not available, sold out, #nogas", it is classified as a "not available" claim. Similarly, if a tweet contains a keyword such as "#opengas, x dollar/gallon, available", it is classified as an "available" claim. We ignore claims that cannot be classified.



@NJGas Valero 120 Watchung Ave. Montclair NJ had 1 car. No lines. :-)

2:08 PM - 8 Nov 2012

#njgas edison costco very small line. Even only no cash

2:18 PM - 8 Nov 2012

In this example, the first tweet can be mapped to an exact address "120 Watchung Ave, Montclair, NJ" which is listed in the ground truth gas locations. However, the second tweet cannot be mapped to an exact station, thus discarded.

Figure 4: Example Gas Locations Reported by Twitter Users

2) Baseline Methods and Parameter Settings: We chose the following four representative truth discovery solutions as the baselines in the evaluation.

- **Voting:** The basic truth discovery algorithm that determines the true value of a measured variable by picking the one with most claims (votes).
- **SSTD:** A state-of-the-art dynamic truth discovery scheme based on an extended Hidden Markov Model [9].
- **TruthFinder (TF):** An iterative truth discovery algorithm that uses a pseudo-probabilistic model [18].
- **EM:** A truth discovery solution based on Maximum Likelihood Estimation that jointly identifies the true values of measured variables and the source reliability [2].

Since none of these methods are designed to handle missing data, we integrate two interpolation techniques (ARMA [43] and K-nearest-neighbour (KNN) [39]) with these baselines to address the missing truth challenge. In particular, ARMA is a temporal interpolation technique that infers current true value based on historical values (discussed in our previous section). KNN, on the other hand, is a spatial interpolation technique that estimates the missing value based on the average values of the neighbors. Here we use physical affinity of gas stations ($\leq$10 miles) to define the neighborhood of measured variables. For batch-based algorithms (i.e., Voting, TD, and EM), we run the algorithms periodically in each sensing cycle. For our scheme, we set maximum lag $Q = 2$, temporal order $P = 3$, $\lambda_1 = 3$ and $\lambda_2 = 1$. These parameters are estimated using a training set of 10 gas stations with full true values. For other baselines with tunable parameters, we find their best settings using the same training set as the RMTF.

3) Evaluation Metrics: We first use the classical metrics for binary classification: *Precision*, *Recall* and *F1-Score*. We did not include *accuracy* as the metric because the collected dataset is very imbalanced: only 17.3% of the measured variables are labeled as negative (i.e., "not available").

Therefore, we introduce two extra metrics specifically designed for imbalanced data, namely Cohen's Kappa (*Kappa*)

[2] http://apollo.cse.nd.edu/datasets.html
[3] http://www.ahcusa.org/ahcs-system.html

Table II: Data Trace Statistics

Collection Period	Nov. 2nd 2012 - Nov. 15th 2012
Number of Tweets	1,511
Number of Sources	1,230
Number of Measured Variables	98
Percentage of Missing Truth	45.03%
Locations of Measured Variables	New Jersey & New York (State), USA
Hashtags	#NJGas, #NYGas, #NYCGas, #queensgas,#nogas, #opengas

and Matthews Correlation Coefficient (*MCC*). MCC returns a value between -1 and 1, where 1 implies a perfect prediction, 0 implies no better than random guess and -1 represents the total disagreement between predicted value and true value. Similarly, Cohen's Kappa also measures the agreement between the true value and predicted value. If the two values are in a complete agreement, then $Kappa = 1$. For all the experiments, we perform one-step ahead inference such that all the historical data are used to infer the values of the current sensing cycle. The performance metrics are calculated based on the inference results of all the sensing cycles.

B. Experiment Results

1) Inference Effectiveness: The results of the inference effectiveness are shown in Table III. We observe our scheme outperforms all baselines in terms of both balanced metrics and imbalanced metrics. For balanced metrics, we found the precision and F-1 scores are high for most of the baselines as shown in Table III. This is due to the fact that the positive samples are dominant in our imbalanced dataset, which cause very high true positives. The imbalanced metrics clearly demonstrate the significant performance gain of RMTF. In particular, compared to the best performing baseline (EM_{KNN}), RMTF has achieved 2.88 times higher Kappa, and 2.49 times higher MCC scores respectively. RMTF achieves such performance gain by explicitly addressing challenges of the missing truth and lagged latent correlation between measured variables as discussed in Section IV. The real-world implication of such performance improvement is that our scheme can accurately identify more stations with gas and minimize the chance of sending the drivers to the stations with no gas.

2) Parameter Sensitivity: We further evaluate the sensitivity of the important parameters λ_1 and λ_2 in the RMTF scheme. First, we study the influence of the spatial correlation λ_1. The results are shown in Figure 5. We set $\lambda_2 = 1$ and vary λ_1 from 0 to 50. We observe that the performance of RMTF without the spatial correlation ($\lambda_1 = 0$) is significantly worse than that with spatial correlation. This demonstrates the importance of incorporating the spatial correlation between measured variables into the framework. We also observe that the performance of RMTF suddenly drops when λ_1 is equal to or greater than 4. The reason is that the temporal disagreement gets overlooked and consequently degrades the performance of RMTF as we increase the weight of the spatial disagreement. Similarly, we study the influence of the temporal correlation by tuning λ_2 from 0 to 50. The result is shown in Figure 6. We also observe that the performance of RMTF significantly drops when the temporal correlation is ignored ($\lambda_2 = 0$). The RMTF reaches its peak performance when $\lambda_2 = 1$. The above results also show that RMTF is less sensitive to the temporal correlation parameter (i.e., λ_2) than the spatial correlation parameter (ie., λ_1).

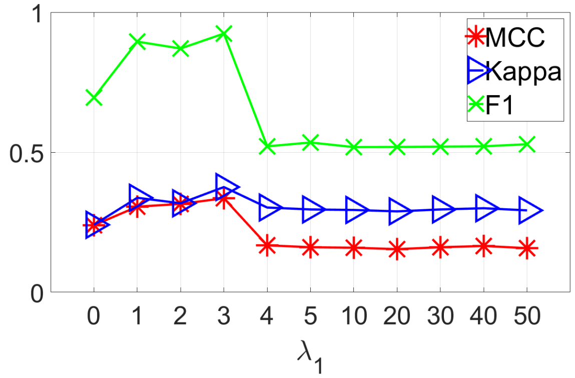


Figure 5: Influence of Spatial Correlation

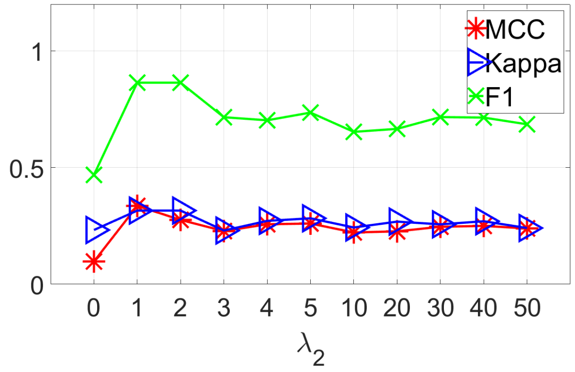


Figure 6: Influence of Temporal Correlation

VI. Limitation and Future Work

Our work has some limitations for future work. The proposed work develops a generic spatiotemporal inference framework to solve dynamic truth discovery problem in social media sensing applications. However, it is likely that spatial or temporal dependencies can be weak or do not exist in some real world applications, which could undermine the

Table III: Evaluation Results

	Kappa	MCC	Precision	Recall	F1-Score
RMTF	**0.3345**	**0.3709**	**0.9730**	**0.8845**	**0.9266**
$\text{Voting}_{\text{ARMA}}$	0.0137	0.0188	0.9371	0.7317	0.8218
$\text{Voting}_{\text{KNN}}$	0.0503	0.0713	0.9456	0.7158	0.8148
$\text{SSTD}_{\text{ARMA}}$	0.1112	0.1261	0.9481	0.8486	0.8956
SSTD_{KNN}	0.0827	0.1052	0.9485	0.7822	0.8574
TF_{ARMA}	0.0525	0.0567	0.9398	0.8705	0.9038
TF_{KNN}	0.0880	0.0967	0.9441	0.8639	0.9022
EM_{ARMA}	0.0890	0.1093	0.9482	0.8021	0.8691
EM_{KNN}	0.1163	0.1491	0.9547	0.7829	0.8603

inference result. To address such problem, we can first test spatiotemporal dependencies of the social media sensing data before we use the RMTF scheme. For example, we can use well-known Durbin–Watson (DW) test [44] or Partial AutoCorrelation Function (PACF) [45] to verify temporal dependency. Similarly, various techniques have been proposed to check the "spatial" correlation among time series data [46]. By leveraging the test results, we can decide whether to incorporate spatial or temporal or both correlations into the RMTF scheme.

Another limitation is that it is possible that the vast majority claims of a measured variable are false, making it difficult to recover the true value of the variable. This may be caused by intentional collusion attack, or unintentional diffusion of a rumor. Unfortunately, this problem has not been well addressed by current truth discovery solutions [17]. One possible solution to this problem is to explicitly incorporate the dependencies between potentially colluded sources into the RMTF scheme. Such dependencies can be estimated based on frequency and timing of social media posts, similarities between posts, and the following/followee relationship between social media users. We can also leverage the propagation pattern of the claims on social media to detect the misinformation spread.

The third limitation of our solution is the linearity assumption of the RMTF model. In particular, the temporal correlation is captured by a linear autoregressive model (i.e., ARMA). Such model lacks expressiveness in modeling complex and non-linear temporal dynamics. To address this limitation, we plan to integrate a well-known statistical approach called "kernel trick" [47] to handle nonlinear models. The "kernel trick" functions will transform the original sensing measurement in such a way that the temporal correlation among measurements is linearized. Another possible approach is to leverage the data-driven models (e.g., Gaussian Process regression) to automatically identify non-linear correlations of time-series data. The authors are actively working on the above directions to address these limitations in their future work.

VII. Conclusion

This paper presents an RMTF scheme to solve the dynamic truth discovery problem with missing truth in online social media sensing applications. In contrast to existing truth discovery solutions that ignore the missing truth issue, the RMTF developed a reliable truth discovery framework to identify the true values of measured variables that have no reported claims by exploring the latent and lagged correlation between variables. We evaluate our solution using a real-world data trace collected from Twitter. The results demonstrate that our solution achieves significant performance gains compared to the state-of-the-art baselines.

Acknowledgement

This research is supported in part by the National Science Foundation under Grant No. CBET-1637251, CNS-1566465 and IIS-1447795, Army Research Office under Grant W911NF-17-1-0409, Google 2018 Faculty Research Award. 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] D. Wang, B. K. Szymanski, T. Abdelzaher, H. Ji, and L. Kaplan, "The age of social sensing," *arXiv preprint arXiv:1801.09116*, 2018.

[2] 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*, Apr. 2012, pp. 233–244.

[3] 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," to appear in Real-Time and Embedded Technology and Applications Symposium (RTAS), 2018 IEEE. IEEE, 2018, accepted.

[4] 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*, Apr. 2014, pp. 35–46.

[5] M. Alger, E. Wilson, T. Gould, R. Whittaker, and N. Radulovic, "Real-time traffic monitoring using mobile phone data," *Online: http://www. maths-in-industry. org/miis/30 Vodafone Pilotentwicklung GmbH*, 2004.

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

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[8] 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*. IEEE, 2016, pp. 1076–1081.

[9] 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*. IEEE, 2017, pp. 966–976.

[10] D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal, "Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications," in *Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference on*. IEEE, 2013, pp. 530–539.

[11] 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*. IEEE, 2017, pp. 57–66.

[12] 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*. IEEE, 2017.

[13] 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*, vol. 5, no. 6, 2012, pp. 550–561.

[14] X. Wang, Q. Z. Sheng, X. S. Fang, L. Yao, X. Xu, and X. Li, "An integrated bayesian approach for effective multi-truth discovery," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 2015, pp. 493–502.

[15] D. Y. Zhang, Q. Li, H. Tong, J. Badilla, Y. Zhang, and D. Wang, "Crowdsourcing-based copyright infringement detection in live video streams," in *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2018*, 2018, accpeted.

[16] C. Huang and D. Wang, "Topic-aware social sensing with arbitrary source dependency graphs," in *Proceedings of the 15th International Conference on Information Processing in Sensor Networks*. IEEE Press, 2016, p. 7.

[17] D. Zhang, D. Wang, N. Vance, Y. Zhang, and S. Mike, "On scalable and robust truth discovery in big data social media sensing applications," *IEEE Transactions on Big Data*, 2018.

[18] 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*, vol. 20, no. 6, pp. 796–808, Jun. 2008.

[19] 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*. IEEE, 2013, pp. 212–223.

[20] X. Wei, J. Sun, and X. Wang, "Dynamic mixture models for multiple time-series." in *Ijcai*, vol. 7, 2007, pp. 2909–2914.

[21] A. Ghasemi and E. S. Sousa, "Opportunistic spectrum access in fading channels through collaborative sensing." *JCM*, vol. 2, no. 2, pp. 71–82, 2007.

[22] S. Ilarri, O. Wolfson, and T. Delot, "Collaborative sensing for urban transportation." *IEEE Data Eng. Bull.*, vol. 37, no. 4, pp. 3–14, 2014.

[23] H.-P. Hsieh, S.-D. Lin, and Y. Zheng, "Inferring air quality for station location recommendation based on urban big data," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 437–446.

[24] D. Y. Zhang, J. Badilla, H. Tong, and D. Wang, "An end-to-end scalable copyright detection system for online video sharing platforms," in *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2018*, 2018, accpeted.

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

[26] Y. Zhang, N. Vance, D. Zhang, and D. Wang, "Optimizing online task allocation for multi-attribute social sensing," in *The 27th International Conference on Computer Communications and Networks (ICCCN 2018)*. IEEE, 2018.

[27] 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)*. ACM, 2016, pp. 167–174.

[28] M. T. Al Amin, T. Abdelzaher, D. Wang, and B. Szymanski, "Crowd-sensing with polarized sources," in *Distributed Computing in Sensor Systems (DCOSS), 2014 IEEE International Conference on*. IEEE, 2014, pp. 67–74.

[29] Y. Zhang, N. Vance, D. Zhang, and D. Wang, "On opinion characterization in social sensing: A multi-view subspace learning approach," to appear in Distributed Computing in Sensor Systems (DCOSS), 2018 International Conference on. IEEE, 2018.

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

[31] 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*, vol. 31, no. 6, pp. 1026–1037, 2013.

[32] D. Wang, L. Kaplan, and T. F. Abdelzaher, "Maximum likelihood analysis of conflicting observations in social sensing," *ACM Transactions on Sensor Networks (ToSN)*, vol. 10, no. 2, p. 30, 2014.

[33] Z. Z. J. Cheng and W. Ng, "Truth discovery in data streams: A single-pass probabilistic approach," in *In Proc. of CIKM*, 2014, pp. 1589–1598.

[34] D. Wang, M. T. Al Amin, T. Abdelzaher, D. Roth, C. R. Voss, L. M. Kaplan, S. Tratz, J. Laoudi, and D. Briesch, "Provenance-assisted classification in social networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 4, pp. 624–637, 2014.

[35] Y. Kamarianakis and P. Prastacos, "Space–time modeling of traffic flow," *Computers & Geosciences*, vol. 31, no. 2, pp. 119–133, 2005.

[36] R. K. Pace, R. Barry, O. W. Gilley, and C. Sirmans, "A method for spatial–temporal forecasting with an application to real estate prices," *International Journal of Forecasting*, vol. 16, no. 2, pp. 229–246, 2000.

[37] D. S. Stoffer, "Estimation and identification of space-time armax models in the presence of missing data," *Journal of the American Statistical Association*, vol. 81, no. 395, pp. 762–772, 1986.

[38] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4597–4605.

[39] L. Wang, D. Zhang, A. Pathak, C. Chen, H. Xiong, D. Yang, and Y. Wang, "Ccs-ta: quality-guaranteed online task allocation in compressive crowdsensing," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015.

[40] M. Umer, L. Kulik, and E. Tanin, "Spatial interpolation in wireless sensor networks: localized algorithms for variogram modeling and kriging," *Geoinformatica*, vol. 14, no. 1, p. 101, 2010.

[41] P. J. Moreno, P. P. Ho, and N. Vasconcelos, "A kullback-leibler divergence based kernel for svm classification in multimedia applications," in *Advances in neural information processing systems*, 2004.

[42] T. Theunissen, "Binary programming and test design," *Psychometrika*, vol. 50, no. 4, pp. 411–420, 1985.

[43] J. S. Huang, "The autoregressive moving average model for spatial analysis," *Australian & New Zealand Journal of Statistics*, vol. 26, no. 2, pp. 169–178, 1984.

[44] J. Durbin and G. S. Watson, "Testing for serial correlation in least squares regression. ii," *Biometrika*, vol. 38, no. 1-2, pp. 159–178, 1951.

[45] F. L. Ramsey *et al.*, "Characterization of the partial autocorrelation function," *The Annals of Statistics*, vol. 2, no. 6, pp. 1296–1301, 1974.

[46] J. P. Buonaccorsi, J. S. Elkinton, S. R. Evans, and A. M. Liebhold, "Measuring and testing for spatial synchrony," *Ecology*, vol. 82, no. 6, pp. 1668–1679, 2001.

[47] M. Martínez-Ramón, J. L. Rojo-Alvarez, G. Camps-Valls, J. Muñoz-Marí, E. Soria-Olivas, A. R. Figueiras-Vidal *et al.*, "Support vector machines for nonlinear kernel arma system identification," *IEEE Transactions on Neural Networks*, vol. 17, no. 6, pp. 1617–1622, 2006.