A Constrained Maximum Likelihood Estimator for Unguided Social Sensing

Huajie Shao*, Shuochao Yao*, Yiran Zhao*, Chao Zhang*,
Jinda Han*, Lance Kaplan†, Lu Su‡, Tarek Abdelzaher*
* University of Illinois at Urbana-Champaign, Urbana, IL 61801
†US Army Research Labs, Adelphi, MD 20783
‡ State University of New York at Buffalo, Buffalo, NY 14260
Email: {hshao5, syao9, zhao97, czhang82, jhan51, zaher} @illinois.edu,
lance.m.kaplan@us.army.mil, lusu@buffalo.edu

***Abstract*—This paper develops a constrained expectation maximization algorithm (CEM) that improves the accuracy of truth estimation in *unguided* social sensing applications. Unguided social sensing refers to the act of leveraging naturally occurring observations on social media as "sensor measurements", when the sources post at will and not in response to specific sensing campaigns or surveys. A key challenge in social sensing, in general, lies in estimating the veracity of reported observations, when the sources reporting these observations are of unknown reliability and their observations themselves cannot be readily verified. This problem is known as *fact-finding*. Unsupervised solutions have been proposed to the fact-finding problem that explore notions of internal data consistency in order to estimate observation veracity. This paper observes that unguided social sensing gives rise to a new (and very simple) constraint that dramatically reduces the space of feasible fact-finding solutions, hence significantly improving the quality of fact-finding results. The constraint relies on a simple approximate test of source independence, applicable to unguided sensing, and incorporates information about the number of *independent* sources of an observation to constrain the posterior estimate of its probability of correctness. Two different approaches are developed to test the independence of sources for purposes of applying this constraint, leading to two flavors of the CEM algorithm, we call CEM and CEM-Jaccard. We show using both simulation and real data sets collected from Twitter that by forcing the algorithm to converge to a solution in which the constraint is satisfied, the quality of solutions is significantly improved.**

***Index Terms*—social networks, truth discovery, constrained expectation maximization (CEM), estimation accuracy**

I. Introduction

This paper proposes a constrained expectation maximization algorithm (CEM) that enhances the quality of truth estimation (i.e., fact-finding) in *unguided* social sensing applications. By *unguided* social sensing, we refer to the act of leveraging naturally occuring observations on social media as "sensor measurements", when the sources post at will and *not in response to specific sensing campaigns or surveys*. This is in contrast to situations, where participating sources are asked (e.g., via a crowdsensing phone app or a survey form) to answer *specific questions*. The main challenge is to fuse information from mutltiple sources to get estimates of the probability of correctness of reported observations [1]. Thus far, a significant number of fact-finding approaches have been proposed in various areas, such as fake news discovery on social media [2]–[4], and crowdsourcing [5]–[7]. We explore the possibility of improving the solutions by appropriately constraining the solution space.

It is generally acknowledged, in past literature, that sources who make truthful observations will usually agree because truth (in the sense of being a state that matches physical reality) is unique. On the other hand, multiple versions of deviation from the truth are possible. For example, if an escape vehicle used in a given robbery was blue, observers who report the color incorrectly may erroneously mention any one of a large variety of other colors. Hence, coincidental agreement among incorrect observations is less likely. This is true unless the sources are somehow non-independent; for example, if one source is copying from another. On social media, such as Twitter, copying behavior can sometimes be directly detected. A retweet of an original tweet is, by definition, a copy. In general, however, copying behavior can occur outside of retweets as well, and as such needs to be detected separately. This is challenging because, after all, truthful sources *should* be correlated because they report the same truth. Hence, when the goal is to determine what's true, how does one distinguish whether agreement on observations is a sign of copying behavior or simply an indicator that the observations are true? This question is at the core of distinguishing truth from rumors.

In this paper, we observe that in *unguided* social sensing, the above distinction can be made with relative ease. Consider observations shared on social media (such as Twitter), where individuals are not reporting in response to a targeted crowdsensing campaign or a specific survey question, but rather reporting whatever they choose. This setting offers a different way of observing non-independence based on an intuition we call the *infinite domain consideration*. To understand the intuition, consider a "take-home" exam, where the instructor wants to detect whether or not students are copying. Let the exam have a very large (in the limit, infinite) number of questions. The students are asked to answer a small finite number of them, which they can choose at will. If two students happened to choose to answer a largely overlapping subset of questions out of the infinite set, it is statistically very likely

978-1-5386-4128-6/18/$31.00 ©2018 IEEE

that they have copied from one another, regardless of whether their answers are in fact correct or not.

Similarly, given all the possible observations that two users in an unguided social sensing scenario can make on the social medium (about anything), if they repeatedly choose the same things to observe, then some copying behavior or a shared influence is suspected. Unlike what the case might be with the exam, it is not our objective to penalize copying. Rather, our objective is to detect evidence of the absence thereof (i.e., lack of an appreciable overlap in observations), which would then suggest independence of the sources in question. If two or more sources, who are deemed independent *by the above test*, agree on an observation, then this particular observation is very likely to be true. In fact, it is likely true *regardless of the reliability of the sources in question* because the probability of making the same mistake by independent sources is small. This consideration implies that the probability that an observation is true can be *higher* than what's implied by source reliability alone.

Interestingly, the mathematical analysis underlying most current fact-finding frameworks does not account for the infinite domain consideration, described above. Rather, the analysis expresses the probability of agreement on an observation as a function of the reliability of sources and whether the observation is true or false. This allows the analysis to infer the latter from the former (often iteratively). However, the analysis does not consider the equivalent of the "number of questions on the test" in our illustrative example. We show that by retrofitting a very simple constraint into the existing maximum likelihood estimation framework, it is possible to account for the infinite domain consideration, thereby significantly improving the results.

In this paper, we develop a novel constrained expectation maximization (CEM) algorithm to assess the correctness of the observed data in *unguided* social sensing. We evaluate two flavors of the proposed algorithm, called CEM and CEM-Jaccard, using both synthetic data and real-world datasets collected from Twitter. Evaluation results demonstrate that both the CEM algorithm and CEM-Jaccard algorithm outperform the baselines of truth-finders in social sensing, and that CEM-Jaccard does better in scenarios, where shared biases within communities increase instances of agreement on the same incorrect observations.

The rest of the paper is organized as follows. We develop the constrained expectation maximization method in Section II. In Section III, we evaluate the estimation accuracy of the proposed algorithms both using simulations and real-world datasets. Section IV summaries the related work on truth discovery. Finally, we conclude the paper in Section V.

II. Constrained Expectation Maximization

Below, we first present general background on expectation maximization (EM) and constrained expectation maximization (CEM) algorithms in Section II-A. Section II-B then develops the mathematical problem formulation for unguided social sensing, offering two flavors of CEM that differ in the used constraints. Section II-C summarizes the resulting algorithm. Finally, Section II-D addresses the challenge with malicious users. As one might remember from the introduction, our observation was that, in unguided social sensing, claims are more likely to be true if made by multiple independent users than what might be suggested by source reliability alone. This, however, opens the door for abuse by malicious users who fool the algorithm into assuming source independence. A solution that significantly limits user ability to abuse the algorithm is presented in Section II-D.

A. Background

The general expectation maximization (EM) algorithm [8], [9] aims to find estimates of some parameter set that maximize the likelihood of collected observations, when those observations depend on both the parameter values to be estimated as well as some latent variables [10]. Given an observed data set $\mathbf{X}$, latent variables $\mathbf{Z}$, and the unknown parameters θ, a log maximum likelihood is defined:

$$\begin{aligned} L(\theta;\mathbf{X}) &= \log \sum_{\mathbf{Z}} p_\theta(\mathbf{X},\mathbf{Z}) \\ &= \log \sum_{\mathbf{Z}} q(\mathbf{Z}|\mathbf{X}) \frac{p_\theta(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z}|\mathbf{X})} \\ &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}|\mathbf{X}) \log \frac{p_\theta(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z}|\mathbf{X})}, \end{aligned} \tag{1}$$

where $q(\mathbf{Z}|\mathbf{X})$ is the posterior of latent variables. The expectation expression on the right-hand side can then be maximized with respect to values of estimated parameters, as well as latent variables, leading the standard EM algorithm to iteratively implement two steps, called the E-step and M-step, to estimate the unknown parameters θ and latent variables, $\mathbf{Z}$. The posterior $q(\mathbf{Z}|\mathbf{X})$ in the E-step is determined by $p_\theta(\mathbf{Z}|\mathbf{X})$ according to KL divergence [10].

Different from the standard EM algorithm, the constrained expectation maximization (CEM) algorithm directly specifies some prior information about the posteriors through some features of the observed data. The posterior $q(\mathbf{Z}|\mathbf{X})$ is not only determined by $p_\theta(\mathbf{Z}|\mathbf{X})$, but also related to the prior information about some features of the observed data. As before, the CEM algorithm is divided into two steps: the E-Step (now with constraint) and the M-Step [11].

- E-step with constraint: This step restricts the posteriors of latent variables, such that $q(\mathbf{Z}|\mathbf{X}) \in \mathcal{Q}(\mathbf{X})$, instead of restricting $p_{\theta^t}(\mathbf{Z}|\mathbf{X})$ directly. Formally,

$$q^{t+1}(\mathbf{Z}|\mathbf{X}) = \underset{q(\mathbf{Z}|\mathbf{X})\in\mathcal{Q}(\mathbf{X})}{\arg\min} \; KL\left(q(\mathbf{Z}|\mathbf{X})||p_{\theta^t}(\mathbf{Z}|\mathbf{X})\right). \tag{2}$$

- M-step: This step is used to estimate the unknown parameters θ given the posteriors. It remains unchanged from the original EM algorithm:

$$\arg\max_{\theta} \; E\left[\sum_{\mathbf{Z}} q^{t+1}(\mathbf{Z}|\mathbf{X}) \log p_\theta(\mathbf{X},\mathbf{Z})\right]. \tag{3}$$

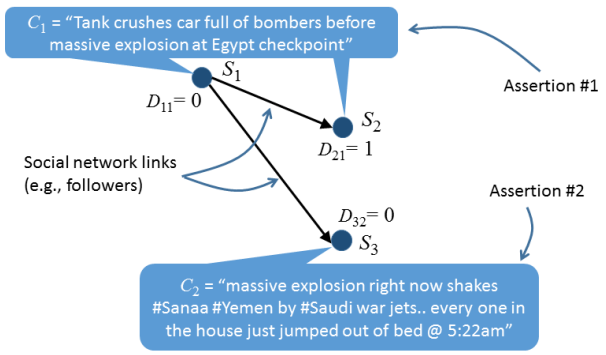


Fig. 1. Illustration of sources are connected by social graphs.

B. Unguided Social Sensing

To use this algorithm for fact-finding, assume that a group of N sources, $\mathcal{S} = \{S_1, S_2, \ldots, S_N\}$, report M assertions. An assertion here refers to a statement that a source makes (e.g., a tweet). In general, multiple sources may agree on the same statement, and multiple statements may be made by the same source. Let C_j denote a Boolean variable associated with the jth assertion to indicate whether it is true or not. We say that $C_j = 1$, if the assertion is *true*. Otherwise, $C_j = 0$, if it is *false*. The act of reporting an assertion by a source is called a *claim* made by that source. When source S_i reports assertion C_j, we say that $S_iC_j = 1$. Otherwise, we say that $S_iC_j = 0$. The elements S_iC_j can thus be thought of as cells of a two dimensional $(N \times M)$ matrix, called the observation matrix, SC. At the risk of abusing the notation, we denote by C_j both the statement of the assertion itself, as well as the Boolean variable denoting its truth value.

On a social medium, such as Twitter, one can also observe the topology of the underlying social network, such as who is following whom, or who is retweeting whom. Hence, sources are connected by a graph, where downstream nodes have a chance to observe claims made by upstream ones, as shown in Fig. 1. When source S_i is a successor of source S_k, there exists an edge from S_k to S_i. We call sources downstream from S_k its *successors*, and call S_k their *ancestor*. Those with no incoming edges are roots. We introduce the indicator D_{ij} to denote whether a certain source S_i has ancestors who made assertion C_j or not. Let $D_{ij} = 1$ denote that some ancestor of S_i made assertion C_j. $D_{ij} = 0$ represents that no ancestor of S_i made assertion C_j. We further define D as the two-dimensional matrix of all such indicators. Hence, if $S_iC_j = 1$ and $D_{ij} = 1$, then source S_i may be repeating observation C_j because an ancestor of theirs made the same assertion. Figure 1 shows indicators D_{ij} for three sources making two assertions. In this case, $D_{11} = 0$ because source S_1 has no ancestors making assertion C_1 = "Tank crushes car full of bombers before massive explosion at Egypt checkpoint". On the other hand, $D_{21} = 1$ because source S_2 has an ancestor who makes assertion C_1. Also, $D_{32} = 0$ because source S_3 has no ancestor who makes assertion C_2.

Inspired by the approach in [12], we define the following parameters to be estimated:

- $a_i = P(S_iC_j = 1|C_j = 1, D_{ij} = 0)$: The probability that source S_i reports C_j, given that assertion C_j is true and no ancestor of S_i previously reported the same assertion.
- $b_i = P(S_iC_j = 1|C_j = 0, D_{ij} = 0)$: The probability that source S_i reports C_j when it is in fact false and no ascestor of S_i previously reported the same assertion.
- $f_i = P(S_iC_j = 1|C_j = 1, D_{ij} = 1)$: The probability that source S_i report C_j, given that assertion C_j is true and some ancestors of S_i previously reported the same assertion.
- $g_i = P(S_iC_j = 1|C_j = 0, D_{ij} = 1)$: The probability of source S_i reports C_j when it is in fact false and some ancestors of S_i previously reported the same assertion.

For our problem of truth discovery, the latent variables are the set, $C = \{C_1, C_2, \ldots, C_M\}$, denoting whether the assertions are true or false. They must be determined based on the observation matrix, SC, denoting which sources made which assertions. Let d denote unknown expected ratio of correct assertions, $P(C_j = 1)$, and $\theta = [d; \forall i : a_i, b_i, f_i, g_i]$ denote the unknown source parameters we need to estimate. Our goal is to estimate the unknown parameters θ together with the truth value of each assertion, C, given the observed data, SC, and the social graph, D. This paper adopts variations of the CEM algorithm to solve the truth discovery problem. The log likelihood function is thus given by:

$$\begin{aligned}\mathcal{L} &= \log P(SC; D, \theta) \\ &= \log\left(\sum_{C\in\{0,1\}} P(SC|C; D, \theta)P(C; D, \theta)\right).\end{aligned} \tag{4}$$

The E-step becomes:

$$\begin{aligned}Q(\theta|\theta^t) = &\sum_{C\in\{0,1\}} P(C|SC; D, \theta^t)\times \\ &\log\Big(P(SC|C; D, \theta)P(C; D, \theta)\Big).\end{aligned} \tag{5}$$

where $P(C|SC; D, \theta^t)$ is the posterior probability of latent variable C.

Since there exists M assertions in set C, Eq. (5), above, could be rewritten as:

$$\begin{aligned}Q(\theta|\theta^t) = &\sum_{j=1}^{M}\sum_{C_j\in\{0,1\}} P(C_j|SC_j; D, \theta^t)\log\Big\{P(C_j; D, \theta) \\ &P(SC_j|C_j; D, \theta)\Big\},\end{aligned} \tag{6}$$

where $P(C_j|SC_j; D, \theta^t)$ could be expressed by:

$$P(C_j|SC_j; D, \theta^t) = \frac{P(SC_j|C_j; D, \theta^t)P(C_j; D, \theta^t)}{\sum_{C_j\in\{0,1\}} P(SC_j|C_j; D, \theta^t)P(C_j; D, \theta^t)}. \tag{7}$$

For the above Eq. (7), SC_j refers to the jth claim made by N sources, so we have:

$$P(SC_j|C_j; D, \theta^t) = \prod_{i=1}^{N} P(S_iC_j|C_j; D, \theta^t), \tag{8}$$

where $P(SC_j|C_j; D, \theta^t)$ is corresponding to the parameters a_i, b_i, f_i, g_i defined above given different D_{ij} and C_j.

We can now hypothesize that, if two sources are independent, the likelihood that they make the same false assertion is very low. The question is: how do we decide that the sources are independent? In this paper, we propose an answer based on the consideration mentioned in the introduction. First, trivially, a source should not be repeating a claim made by one of its ancestors, since this would not be independent. Second, the sources in question should not have a history of coinciding assertions. Let the set of sources that made an assertion C_j and for whom the above constraints are met be N_j, where $|N_j| = n_j$.

Consider assertion C_j that represents some statement about the state of the world. In the simplest case, the state has only two possible values (one of which must be correct and one of which is wrong). Let n_j independent sources make the same assertion. Either the sources agreed on this assertion because it is correct, or they agreed because they accidentally made the same error. If the sources are guessing randomly, after the first source guesses, the probability that the rest make the same error is defined by λ^{n_j-1}, where $0 < \lambda < 1$. This probability decreases with n_j. As a heuristic, we take $1 - \lambda^{n_j-1}$ to be the probability that sources are not guessing randomly (but rather agree on a real observation of the truth). It can be easily seen that this probability will be higher if the physical state in question had more than two possible values, because the probability of coincidental agreement will be even lower. Therefore, when agreement among n_j *independent* sources occurs, we constrain the probability of correctness $q(C_j|SC_j) \geq 1 - \lambda^{n_j-1}$.

It remains to decide on whether sources are independent. We consider two versions of the algorithm. In the first (plain CEM), independence is assumed simply if each pair of sources in N_j do not have an ancestor/descendant relation. In the second, we also ensure that they have substantially different histories of claims. Let S_iC and S_kC denote the set of past assertions reported by S_i and S_k, respectively. We define:

$$J(S_i, S_k) = \frac{|S_iC \cap S_kC|}{|S_iC \cup S_kC|}. \tag{9}$$

as a measure of similarity based on the Jaccard distance. If $J(S_i, S_k) > threshold$, we think of S_i and S_k as non-independent and count them as one source for purposes of computing n_j. Correspondingly, we call this approach CEM-Jaccard in this paper.

Hence, the E-step for our truth discovery problem is given by:

$$\begin{aligned} &q^{t+1}(C_j|SC_j) = \underset{q(C_j|SC_j)}{\arg\min}\ KL\left(q(C_j|SC_j)||P(C_j|SC_j; D, \theta^t)\right), \\ &\text{s.t., } \forall_j: \quad 1 - \lambda^{n_j-1} \leq q(C_j|SC_j) \leq 1. \end{aligned} \tag{10}$$

We call the above constraint, the *probability boosting constraint*, since its main effect is to boost the probability of correctness of some claims, compared to the unconstrained version. Specifically, when the posterior probability computed by the KL divergence is less than the lower bound $1 - \lambda^{n_j-1}$, it will be boosted to this bound.

Next, the M-step is used to maximize the $Q(\theta|\theta^t)$ to estimate the unknown parameter θ given the posterior of latent variable $q^t(C_j|SC_j)$, yielding

$$\theta^{t+1} = \arg\max Q(\theta|\theta^t). \tag{11}$$

The optimal estimation of the unknown parameters θ can be obtained through multiple iterations of the derivative $\frac{\partial Q(\theta|\theta^t)}{a_i} = 0$, $\frac{\partial Q(\theta|\theta^t)}{b_i} = 0$, $\frac{\partial Q(\theta|\theta^t)}{f_i} = 0$, $\frac{\partial Q(\theta|\theta^t)}{g_i} = 0$, $\frac{\partial Q(\theta|\theta^t)}{d} = 0$. So we can have

$$a_i^{t+1} = \frac{\sum_{C_j \in S_iC_j^{D_0}=1} q^t(C_j|SC_j)}{\sum_{C_j \in S_iC_j^{D_0}} q^t(C_j|SC_j)}, \tag{12a}$$

$$b_i^{t+1} = \frac{\sum_{C_j \in S_iC_j^{D_0}=1} \left(1 - q^t(C_j|SC_j)\right)}{\sum_{C_j \in S_iC_j^{D_0}} \left(1 - q^t(C_j|SC_j)\right)}, \tag{12b}$$

$$f_i^{t+1} = \frac{\sum_{C_j \in S_iC_j^{D_1}=1} q^t(C_j|SC_j)}{\sum_{C_j \in S_iC_j^{D_1}} q^t(C_j|SC_j)}, \tag{12c}$$

$$g_i^{t+1} = \frac{\sum_{C_j \in S_iC_j^{D_1}=1} \left(1 - q^t(C_j|SC_j)\right)}{\sum_{C_j \in S_iC_j^{D_1}} \left(1 - q^t(C_j|SC_j)\right)}, \tag{12d}$$

$$d^{t+1} = \frac{\sum_{j=1}^{M} q^t(C_j|SC_j)}{M}, \tag{12e}$$

where $S_iC_j^{D_0} = \{S_iC_j : \forall S_iC_j \in SC \& D_{ij} = 0\}$; and $S_iC_j^{D_1} = \{S_iC_j : \forall S_iC_j \in SC \& D_{ij} = 1\}$. M is the total number of assertions reported by sources.

C. Algorithm

In this subsection, we summarize the CEM algorithm and the CEM-Jaccard algorithm for truth discovery problem above.

In Algorithm 1, we first compute the posteriors of latent variables through the E-step and then calculate the unknown parameters in the M-step using the maximum likelihood method. Lines 6 and 7 apply the probability boosting constraint. After the parameters converge, we classify the assertions based on its estimated truth values. If the truth value of an assertion is equal or greater than 0.5, the assertion is evaluated to be true, $\hat{C}_j = 1$; otherwise it is evaluated to be false, $\hat{C}_j = 0$. Note that, the difference between CEM and CEM-Jaccard is in how n_j is computed. In CEM n_j denotes the number of sources who made the same assertion, C_j, who do not have an ancestor/descendant relation. In CEM-Jaccard, it is the number of sources who meet the above condition and also do not have a significant overlap in past tweets, according to the Jaccard distance as expressed in Equation (9).

D. Malicious Users

In the above discussion, we have not considered malicious users who purposely try to subvert the algorithm by posting assertions in a way that falsely increases their credibility. In general, this problem has no solution. A source can always engender trust by posting a number of claims that are true

Algorithm 1 CEM/CEM-Jaccard algorithm for Truth Finding.

Input: initialize θ with random probability, n_j
Output: Classification results: $\hat{C}_j$, θ
while θ^t does not converge **do**
  for $j = 1 : M$ **do**
    Compute posterior $q^{t+1}(C_j|SC_j)$ based on Eq. (10)
    if $q^{t+1}(C_j|SC_j) < 1 - \lambda^{n_j - 1}$ **then**
      $q^{t+1}(C_j|SC_j) = 1 - \lambda^{n_j - 1}$
    else
      $q^{t+1}(C_j|SC_j) = P(C_j|SC_j; \theta^t)$
  end for
  for $i = 1 : N$ **do**
    Compute $a_i^{t+1}, b_i^{t+1}, f_i^{t+1}, g_i^{t+1}, d^{t+1}$ based on Eq. (12)
  end for
  Update $\theta^{t+1} = \theta^t$
  $t = t + 1$
end while
for $j = 1 : M$ **do**
  if $q^t(C_j|SC_j) \geq 0.5$ **then**
    $\hat{C}_j$ = 1 (true)
  else
    $\hat{C}_j$ = 0 (false)
end for
Return classification results $\hat{C}_j$

so its estimated reliability increases then positing a lie that would be considered reliable truth. However, the approaches for assessing source independence, described in this paper, introduce a new vulnerability. Namely, two or more sources with no ancestor/descendant relation in the social network graph can now agree to post the same claim. This may cause CEM to consider this claim correct because the new constraint will regard the probability of coincidental agreement among the independent sources unlikely. CEM-Jaccard can be fooled as well if the sources also fabricate a number of additional claims that are different, so that the Jaccard similarity score between their claim histories is lowered to where they are considered independent. To avoid significant degradation because of these problems, we simply limit the use of the new constraint. That's to say, we limit a single source to be a beneficiary of the constraint at most W times per a certain interval of time. Hence, when enforcement of the constraint boosts the correctness probability of a claim made by n_j sources (that are deemed independent by our algorithm), we increment a per-source counter. Once the counter of some source reaches W, no other claims of that source are subjected to the (probability boosting) constraint for the remainder of the current window. Setting W involves a trade-off. A value that is too small will essentially mean that the new constraint is not sufficiently exercised, which reduces the benefit from the new approach. On the other hand, a value that is too large allows frequent misuse.

This is not unlike the rationale for limiting ATM withdrawals to some maximum amount per defined period (e.g., per day). While the bank cannot prevent someone from stealing a debit card, they can limit the damage by limiting the benefit received from unauthorized use of the card. As with the choice of W, a balance must be maintained between allowing legitimate users to benefit from their debit cards and limiting the damage caused by unauthorized transactions. Our evaluation results on Twitter show that setting W to (approximately) once every 36 hours offers a good compromise between performance benefits and risk. A node will be able to use/abuse the constraint less than once a day, which limits damage. At the same time, the evaluation shows, most of the benefit is achieved.

III. Performance Evaluation

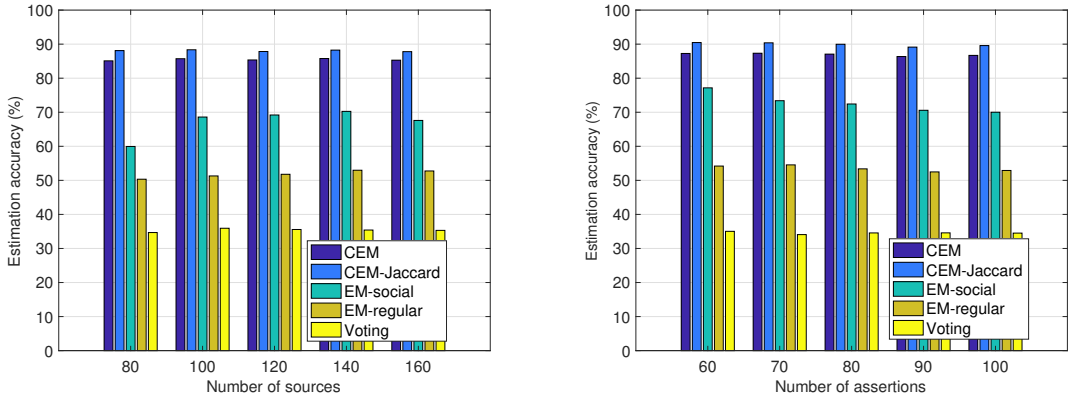
In this section, we carry out extensive experiments to evaluate the performance of the proposed CEM algorithm and CEM-Jaccard algorithm for truth discovery using both synthetic data and real-world datasets collected from Twitter. In addition, we explore limiting the number of times the probability boosting constraint can be used in a given time window for each source to impede malicious sources.

A. Simulation

In this subsection, we use simulations to evaluate the estimation accuracy of the proposed CEM algorithm, CEM-Jaccard algorithm, and the baselines below.

- **CEM**: Our proposed algorithm in this paper. It incorporates prior information on the number of independent sources, derived from the topology of the social graph, to constrain the probability of latent variables.
- **CEM-Jaccard**: Another proposed algorithm in this paper. It computes the number of independent sources, n_j, not only based on the topology of the social graph, but also, in part, based on the Jaccard distance between claim histories of sources.
- **EM-social**: This algorithm was proposed in IPSN'14 [13]. It uses the general (unconstrained) EM algorithm to estimate the truth values of assertions given the social graph.
- **EM-regular**: This algorithm was proposed in IPSN'12 [13] and uses the general EM algorithm to assess the correctness of assertions, *without* considering the social graph (i.e., it assumes that all sources are independent).
- **Voting**: This method is a naive method that estimates veracity of assertions based on the number of sources who claimed them.

To generate synthetic data, we consider a situation where N sources are connected by a heavy-tailed social graph [14] and together make a total of M different assertions. For each assertion, the ground truth validity is determined by flipping a weighted coin, so that it is true with probability P_t. Nodes in the social graph pick assertions to report, such that they report "true assertions" with probability p_i^{indT} and "false assertions" with probability p_i^{indF}, respectively. These values, therefore, determine the predisposition of nodes for making true and false statements, respectively. Their children then have a probability p_i^{depT} to follow their ancestors in reporting a true assertion and probability p_i^{depF} to follow their ancestors in reporting a false assertion. These probabilities thus simulate gossiping behavior.



(a) # of sources (b) # of assertions

Fig. 2. Comparison of Estimation Accuracy for various algorithm

In addition, we use r^T as the probability that a true assertion is reported by additional sources with no ancestor/descendant relation and r^F as the probability that a false assertion is reported by additional sources with no ancestor/descendant relation. The former simulates multiple independent observations of the truth, whereas the latter simulates impact of bias, where multiple nodes originate the same false information out of a shared incorrect belief.

In our simulation, to compare the estimation accuracy of our proposed CEM algorithm with the baselines, we choose simulation parameters based on real-world datasets collected from Twitter. In general, we find the parameters $r^T \in [0.4, 0.75]$ and $r^F \in [0.1, 0.3]$ in real-world datasets. We also find that about 65% to 85% of all assertions are true. So we set up $r^T = 0.65$, $r^F = 0.15$ and $P_t = 0.7$. In addition, we take $N = 100, M = 80$, and $threshold = 0.5$ as the default values, except where mentioned otherwise in the experiments. We set up parameter $\lambda = 0.5$ as it is one of the optimal choices for our experiments. Since the source-claim matrices are sparse in the real-world datasets, we set up the following parameters to emulate similar SC matrices. For a true assertion, reliable sources with no ancestor/descendant relation report it with $p_i^{indT} \in [0.1, 0.5]$ while unreliable sources with no ancestor/descendant relation report it with $p_i^{indT} \in [0.01, 0.05]$. Their children would follow them with $p_i^{depT} \in [0.7, 0.9]$. For a false assertion, it has r^F probability to be reported by multiple sources with no ancestor/descendant relation. These sources make the same assertions with $p_i^{indF} \in [0.8, 0.9]$. In addition, the reliable sources with no ancestor/descendant relation reports it with probability $p_i^{indF} \in [0.005, 0.01]$. Their children would follow the corresponding ancestors with $p_i^{depF} \in [0.7, 0.9]$.

In the first experiment, we evaluate the estimation accuracy of various approaches given a different number of sources N and a different number of assertions M, respectively. Our results are obtained by averaging over 50 instantiations of the source claim matrices. The simulation results are illustrated in Fig. 2.

Fig. 2(a) shows the comparison of the estimation accuracy for different methods as the number of sources, N, varies from 80 to 160. From this figure, we can observe that the estimation accuracy of the CEM-Jaccard is higher than the CEM algorithm. This is because the CEM-Jaccard algorithm checks for the non-independent sources who have substantial histories of claims that they make together. In addition, both CEM algorithm and CEM-Jaccard algorithm outperform the baselines. We can also observe from Fig. 2(a) that the different algorithms seem fairly invariant to changes in the number of sources.

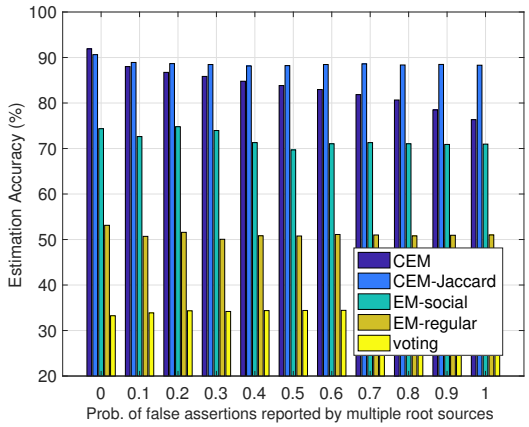


Fig. 3. r^F of false assertions reported by multiple root sources.

Fig. 2(b) compares the estimation accuracy of the above different algorithms for different numbers of assertions. We change M from 60 to 100 with the other parameters kept unchanged. We can observe that both the CEM algorithm and the CEM-Jaccard algorithm outperform the baselines. Likewise, we can observe that the estimation accuracy of the CEM-Jaccard algorithm is a little higher than the CEM.

Next, we explore the effect of bias. Bias in a community affects the probability of reporting false assertions by multiple sources that are not necessarily connected by a social graph and hence will be deemed independent by CEM. Since biased sources agree on views, they will often post the same assertions. To explore this effect, we change the probability,

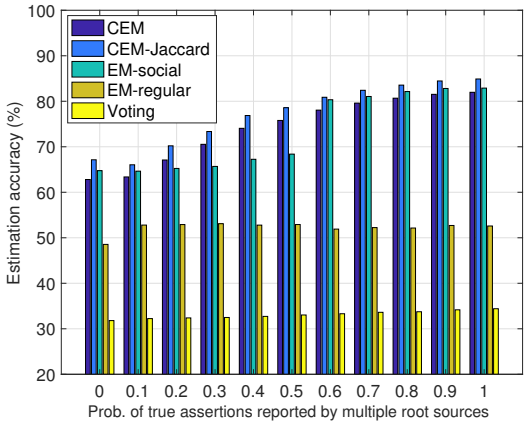


Fig. 4. r^T of true assertions reported by multiple root sources.

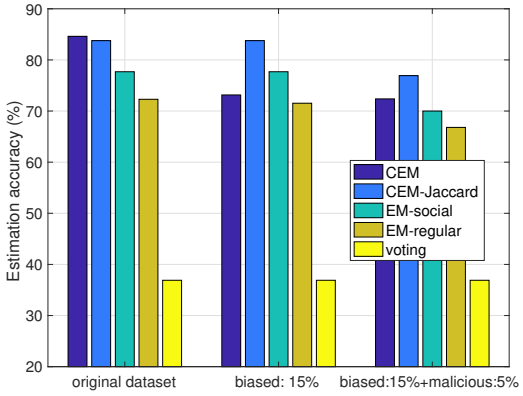


Fig. 5. Comparison of accuracy for different algorithms (Mosul dataset).

r^F, from 0 to 1, while keeping other parameters unchanged. Fig. 3 illustrates that the estimation accuracy of the proposed CEM method will drop with the increase of false assertions reported by multiple sources. But the CEM-Jaccard algorithm still has a higher estimation accuracy than the baselines. This is because the original CEM algorithm constrains the probability to be true just based on ancestor/descendant relation. However, the CEM-Jaccard algorithm checks claim history to assess independence. If sources often coincide in their posted claims, they are no longer deemed independent, and the probability boosting constraint is not used. In addition, when $r^F = 0$, the CEM-algorithm has a little higher estimation accuracy than the CEM-Jaccard, since the CEM-Jaccard may occasionally fail to recognize source independence based on history similarity that is accidental. Based on the evaluation results, we can conclude that the CEM-Jaccard is more robust than the CEM algorithm, especially in the presence of biased communities of sources.

In Fig. 4, we compare the estimation accuracy of various methods for different probabilities of true assertions reported by multiple sources with no ancestor/descendant relation. Let the probability r^T vary from 0 to 1. From Fig. 4, we can observe that the CEM and CEM-Jaccard algorithms beat the baselines as r^T varies from 0.2 to 1. When r^T is less than 0.2, the estimation accuracy of CEM is a little lower than the EM-social since some false assertions (r^F = 0.15) are enforced with constraints. In addition, the estimation accuracy for the EM algorithms gradually increases with r^T. The reason is that they can better exploit corroboration by independent sources. Similarly, we can see the estimation accuracy of the CEM algorithm is a little lower than the CEM-Jaccard method.

B. Empirical Evaluation

In this subsection, we evaluate the performance of the proposed CEM algorithm and CEM-Jaccard algorithm on real-world datasets collected from Twitter.

We collect the real-world datasets from Twitter through their search API that allows one to collect tweets that match keywords. In this study, we collect two real-world data traces related to the *Mosul battle of ISIS in Iraq* in 2017 and *Presidential Campaign of Donald Trump* in 2016 from Twitter. In order to reduce the data sparsity of claims made by various sources, we extract sources that make at least two tweets during the observation period, and use only those sources and their tweets. In our experiments, human graders labelled the ground truth of each assertion (tweet) by investigating the events in question after the fact. Graders mark the assertions as "True", "False" or "Opinion" based on the following rule:

- *True*: Tweets describe physical events that have been verified as true by the grader.
- *False*: Tweets describing events that are false, according to the grader.
- *Opinion*: Subjective comments made by sources like "I love Super Bowl of this year" and "Lady Gaga should not be invited to Super Bowl".

The estimation accuracy is defined as the ratio of true tweets to the total tweets that a particular algorithm believed (i.e., deemed true): #True/(#False + #True + #Opinion).

Next, we implement the different algorithms and baselines to evaluate accuracy based on the two real-world datasets above. For the first empirical experiment, we use the real-world dataset, *Mosul battle of ISIS in Iraq*. Due to grading time limitations, we grade only the most popular 130 assertions. Fig. 5 compares the estimation accuracy for the different methods for the following three scenarios:

- *Original dataset*: The real-world dataset from Twitter.
- *Biased: 15%*: 15% of false assertions from the 130 grades assertions in the real-world dataset are artificially replicated so that they are reported by multiple sources with no ancestor/descendant relation.
- *Biased:15% + Malicious:5%*: We artificially add malicious sources to report 5% of false assertions in the real-world dataset as well. Unlike simply biased sources, malicious sources actively try to fool our algorithm. Hence, they come in groups of 2 or more. Each group makes one common (false) assertion, then each source in the group makes 5 random other assertions in order to fool Jaccard.

From this figure, we can see that the CEM and CEM-Jaccard have higher estimation accuracy than the baselines for the

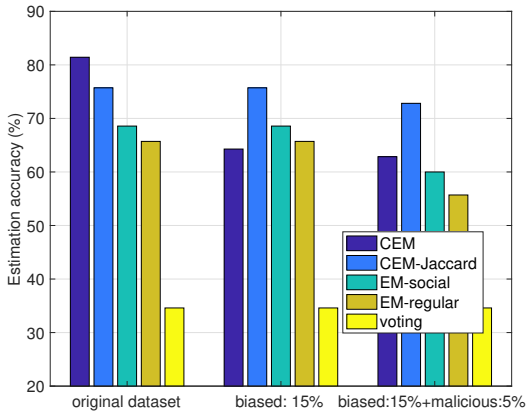


Fig. 6. Comparison of accuracy for different algorithms (Trump dataset).

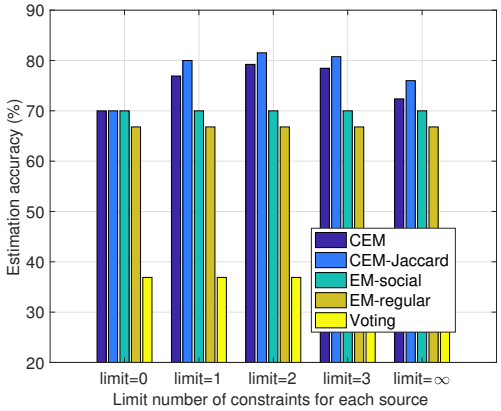


Fig. 7. Limit use of constraints for each source (Mosul dataset).

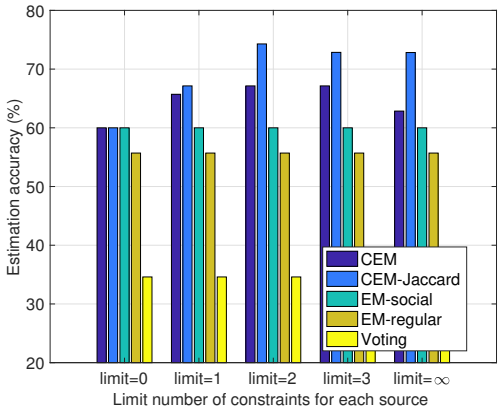


Fig. 8. Limit use of constraints for each source (Trump dataset).

original dataset. This is because about 28% of true assertions reported by multiple sources with no ancestor/descendant relation are enforced with constraints. Note that, in the absence of biased or malicious sources, the estimation accuracy of the CEM-Jaccard is little lower than the CEM, because it removes some constraints unnecessarily when truly independent sources coincidentally make several similar claims. In addition, we can see from Fig. 5 that the estimation accuracy of the CEM seems worse than EM-social for *Biased: 15%*, but the CEM-Jaccard algorithm has a higher estimation accuracy than both. Finally, as multiple malicious sources report false assertions to fool the CEM-Jacard, the estimation accuracy of all methods is decreased in *Biased:15% + Malicious:5%*. Again, CEM-Jaccard does better than the rest.

In the second experiment, the real-world dataset *Presidential Campaign of Donald Trump* is used to compare the estimation performance of different algorithms. We choose 390 assertions as the input of the different algorithms and then choose the 70 most popular assertions to compare their estimation accuracy for the three scenarios as above. As shown in Fig. 6, we can see that the proposed CEM algorithm and the CEM-Jaccard algorithm outperform the baselines for the original dataset. When the ratio of false assertions reported by multiple (biased) sources increases, the estimation accuracy of the CEM algorithm is decreased, whereas CEM-Jaccard is more robust. Finally, we can see the estimation accuracy of the proposed algorithms will be reduced when malicious sources are present. CEM-Jaccard still does the best in this case.

Based on the above evaluation, we conclude that the proposed CEM and CEM-Jaccard methods can achieve better estimation accuracy than the baselines and that CEM-Jaccard is more robust than the CEM algorithm with respect to source bias that results in agreement on many claims by nodes without an actual ancestor/descendant relation in the social graph.

C. Limit of use of constraints

In this subsection, we use the two real-world datasets above to explore how to limit the use of constraints per time interval to impede malicious sources.

As discussed in Section II-D, the proposed CEM algorithm is vulnerable to abuse by malicious sources who may fool it to regard them as independent and hence boost the probability of correctness of claims they agree on. To avoid the resulting degradation in estimation accuracy, we limit the use of constraints for each source. We change the limit for each source from 0 to ∞, applied over a three day window. The evaluation results are as shown in Fig. 7. Note that, when the limit is 0, no constraints are used and the algorithm defaults to EM social. In contrast, when the limit = ∞, constraints are used whenever the algorithm decides that the sources are independent. From Fig. 7, we can see that the estimation accuracy of the CEM and CEM-Jaccard algorithms is higher when $limit = 2$ than the EM-social. Their estimation accuracy is then decreased when increasing the use of constraints. Also, the *Trump dataset* has similar results to the *Mosul dataset* as illustrated in Fig. 8. Thus, limiting the use of constraints for each source to at most two times within a three day interval is best for our data sets. In general the problem of proper limiting deserves further investigation, and will be a topic of expansion for a journal version of this paper.

IV. Related Work

In recent years, much attention was paid to truth discovery from social media. Yin et al [15] developed a TruthFinder that

uses heuristic methods to iteratively estimate the correctness of conflicting data from different websites. Wang et al. [2] proposed an expectation maximization (EM) algorithm to evaluate the correctness of observations in socal sensing. Wang et al. [13] later introduced a source depency model that improved the estimation of reliability of sources and the veracity of assertions by accountind for correlated errors (i.e., rumors) that spread along source dependency chains.

Other researches derived error bounds on reliability. Xiao et al. [16] incorporated source bias into a randomized Gaussian mixture model and built a maximum likelihood estimate (MLE) model for truth discovery. They further derived the theoretical error bounds for population-based and sample-based MLE. However, they assumed that the sources are independent and one will not influence another. Yao et al. [12] derived a fusion error bound given a source dependency model to improve the quality of truth discovery. They calculated the expectation of estimation error and compared several source dependency models.

Some researchers also used supervised learning for truth discovery in social networks. For instance, Castillo et al. [17] assessed the credibility of news on Twitter based on different features from tweets. In reality, we may use deep learning [18] to improve estimation accuracy from various features. However, the drawback of supervised learning is that a lot of data should be labelled.

In addition, some researchers take physical locations into account to detect local events with the help of geo-tags [19], [20]. However, it is often difficult to obtain the real locations of users due to users' privacy protections on social media.

This paper adopted a constrained maximum likelihood estimator to improve the ground truth estimation accuracy in social sensing.

V. Conclusions

This paper developed a novel constrained expectation maximization (CEM) algorithm to improve the accuracy of truth estimation in unguided social sensing. This algorithm incorporates a constraint that boosts the probability of claim correctness as a function of the number of independent sources making the claim. Different from prior EM algorithms, we formulated the E-step as a constrained optimization problem. Two different approaches were proposed to test the independence of sources for purposes of applying this constraint, leading to two flavors of the CEM algorithm, called CEM and CEM-Jaccard. Finally, we evaluated the performance of the CEM and CEM-Jaccard algorithms using both synthetic data and real-world datasets from Twitter. The evaluation results demonstrated that forcing the algorithm to converge to a solution in which the constraint is satisfied, the quality of solutions is significantly improved. In addition, the CEM-Jaccard algorithm is more robust than the CEM algorithm.

Acknowledgements

Research reported in this paper was sponsored in part by the Army Research Laboratory under Cooperative Agreements W911NF-09-2-0053 and W911NF-17-2-0196, in part by DARPA under award W911NF-17-C-0099, and in part by NSF under grants CNS 16-18627 and CNS 13-20209. 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 Laboratory, DARPA, NSF, 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] Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, and J. Han, "On the discovery of evolving truth," in *Proceedings of the 21th ACM SIGKDD*, 2015, pp. 675–684.

[2] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher, "On truth discovery in social sensing: A maximum likelihood estimation approach," in *Proceedings of the 11th IPSN*, 2012, pp. 233–244.

[3] S. Yao, M. T. Amin, L. Su, S. Hu, S. Li, S. Wang, Y. Zhao, T. Abdelzaher, L. Kaplan, C. Aggarwal *et al.*, "Recursive ground truth estimator for social data streams," in *Proceedings of 15th IPSN,*, 2016, pp. 1–12.

[4] H. Shao, S. Wang, S. Li, S. Yao, Y. Zhao, T. Amin, T. Abdelzaher, and L. Kaplan, "Optimizing source selection in social sensing in the presence of influence graphs," in *Proceedings of ICDCS*, 2017, pp. 1157–1167.

[5] H. Jin, L. Su, and K. Nahrstedt, "Theseus: Incentivizing truth discovery in mobile crowd sensing systems," in *Proceedings of MobiHoc 2017*, July, 2017.

[6] R. W. Ouyang, M. Srivastava, A. Toniolo, and T. J. Norman, "Truth discovery in crowdsourced detection of spatial events," *IEEE Trans. on Knowledge and Data Engineering*, vol. 28, no. 4, pp. 1047–1060, 2016.

[7] H. Jin, L. Su, D. Chen, K. Nahrstedt, and J. Xu, "Quality of information aware incentive mechanisms for mobile crowd sensing systems," in *Proceedings of the 16th Mobihoc*, 2015, pp. 167–176.

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

[9] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[10] J. Graça, K. Ganchev, and B. Taskar, "Expectation maximization and posterior constraints." in *Proceedings of NIPS*, vol. 20, 2007, pp. 569–576.

[11] K. Ganchev, J. Gillenwater, B. Taskar *et al.*, "Posterior regularization for structured latent variable models," *Journal of Machine Learning Research*, vol. 11, no. Jul, pp. 2001–2049, 2010.

[12] S. Yao, S. Hu, S. Li, Y. Zhao, L. Su, L. Kaplan, A. Yener, and T. Abdelzaher, "On source dependency models for reliable social sensing: Algorithms and fundamental error bounds," in *Proceedings of the 36th ICDCS*, 2016, pp. 467–476.

[13] D. Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. C. Aggarwal, R. Ganti *et al.*, "Using humans as sensors: An estimation-theoretic perspective," in *Proceedings of IPSN*, 2014, pp. 35–46.

[14] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *Proceedings of the eleventh ACM SIGKDD*, 2005, pp. 177–187.

[15] 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, 2008.

[16] H. Xiao, J. Gao, Z. Wang, S. Wang, L. Su, and H. Liu, "A truth discovery approach with theoretical guarantee," in *Proceedings of SIGKDD*, 2016.

[17] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of WWW*, 2011, pp. 675–684.

[18] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "Deepsense: A unified deep learning framework for time-series mobile sensing data processing," in *Proceedings of the 26th WWW*, 2017, pp. 351–360.

[19] C. Zhang, G. Zhou, Q. Yuan, H. Zhuang, Y. Zheng, L. Kaplan, S. Wang, and J. Han, "Geoburst: Real-time local event detection in geo-tagged tweet streams," in *Proceedings of the 39th SIGIR*, 2016, pp. 513–522.

[20] C. Zhang, K. Zhang, Q. Yuan, H. Peng, Y. Zheng, T. Hanratty, S. Wang, and J. Han, "Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning," in *Proceedings of WWW*, 2017, pp. 361–370.