Truth Discovery With Multi-Modal Data in Social Sensing

Huajie Shao, Dachun Sun, Shuochao Yao, Lu Su, Zhibo Wang, Dongxin Liu, Shengzhong Liu, Lance Kaplan, *Fellow, IEEE*, and Tarek Abdelzaher, *Senior Member, IEEE*

Abstract—This article proposes unsupervised truth-finding algorithms that combine consideration of multi-modal content features with analysis of propagation patterns to evaluate the veracity of observations in social sensing applications. A key social sensing challenge is to develop effective algorithms for estimating both the reliability of sources and the veracity of their observations without prior knowledge. In contrast to prior solutions that use labeled examples to learn content features that are correlated with veracity, our approach is entirely unsupervised. Hence, given no prior training data, we jointly learn the importance of different content features together with the veracity of observations using propagation patterns as an indicator of perceived content reliability. A novel penalized expectation maximization (PEM) algorithm is proposed to improve the quality of estimation results for observations bolstered by multiple features. In addition, we develop a constrained expectation maximum likelihood with multiple features (CEM-MultiF) that introduces a novel constraint to boost the probability of correctness of some claims. Finally, we evaluate the performance of the proposed algorithms, called EM-Multi, CEM-Multi and PEM-MultiF, respectively, on real-world data sets collected from Twitter. The evaluation results demonstrate that the proposed algorithms outperform the existing fact-finding approaches, and offer tunable knobs for controlling robustness/performance trade-offs in the presence of malicious sources.

Index Terms—Social networks, truth discovery, penalized expectation maximization (PEM), multi-modal data, estimation accuracy

1 Introduction

This paper describes novel unsupervised truth-finding algorithms that use multi-modal content features from microblogs to improve the estimation accuracy of truth discovery in social sensing applications. Social sensing refers to the use of sources on social media as "sensors" reporting observations about the physical world. The main challenge lies in developing effective models and algorithms to jointly determine both (i) the correctness of claims and (ii) the reliability of sources, given neither in advance [1]. In the past few years, researchers have proposed a number of truth-finding approaches in various contexts, such as fake news discovery on social media [2], [3], [4], graph knowledge [5], textual pattern discovery [6], [7], [8] and crowdsourcing [9], [10], [11], [12], [13]. This paper develops an unsupervised approach that is novel in that it jointly (i) estimates the veracity of observations and (ii) learns the significance of different veracity indicators, called *corroborating features*, or simply features. It can be perceived as a generalization of prior unsupervised techniques, where the key veracity indicator considered was the reliability of the source. In also extends supervised techniques that considered additional features but required prior training. Specifically, it does not need prior learning to understand the weight of different features used in veracity estimation. Evaluation shows that it improves the estimation accuracy of truth discovery compared to the state of the art. Robustness to malicious sources is also considered.

In prior work, researchers developed maximum likelihood algorithms for truth discovery that attribute true/false binary (truth) values to users' observations [4]. However, as we have shown in earlier work, they do not properly account for the probability of coincidental agreement among independent users [14]. When the same assertion is reported by multiple independent users without copying, its likelihood of being true tends to be statistically high. Motivated by this observation, the authors [14] recently proposed a constrained maximum likelihood (CEM) algorithm that incorporates prior information on the number of independent sources to refine the probability of latent truth variables. This approach boosts estimation accuracy when an assertion is reported by multiple independent sources, but it does not take advantage of other content features. In reality, many posts today corroborate their content by including features such as images or URLs. We use the presence or absence of such items as additional corroborating evidence in this paper. Importantly, we do not actually check that the items (such as images or URLs) indeed corroborate the content. However, our approach does consider the number of sources that propagate the

- *Huajie Shao, Dachun Sun, Shuochao Yao, Dongxin Liu, Shengzhong Liu, and Tarek Abdelzaher are with the Department of Computer Science, the University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA. E-mail: {hshao5, dsun18, syao9, dongxin3, sl29, zaher}@illinois.edu.*
- *Lu Su is with the Department of Computer Science and Engineering, the State University of New York at Buffalo, Buffalo, NY 14260 USA. E-mail: lusu@buffalo.edu.*
- *Zhibo Wang is with the School of National Cybersecurity, the School of Computer, Wuhan University, Wuhan 430072, China. E-mail: zbwang@whu.edu.cn.*
- *Lance Kaplan is with US Army Research Labs, Adelphi, MD 20783 USA. E-mail: lance.m.kaplan.civ@mail.mil.*

Manuscript received 1 Nov. 2019; revised 23 June 2020; accepted 5 July 2020. Date of publication 10 July 2020; date of current version 8 Sept. 2021.
(Corresponding author: Huajie Shao.)
Recommended for acceptance by F. Cappello.
Digital Object Identifier no. 10.1109/TC.2020.3008561

observation. The idea is that, in general, an observation will not propagate well on the medium if it includes images or URLs that are clearly irrelevant. Hence, the presence of such features *together with an appropriate propagation pattern* should increase the likelihood of correctness of the statements made by sources. Take a car accident as an example. If many sources on Twitter reported that "there was a car accident on the highway I57 in Urbana, Illinois" with supporting images and/or news URLs, the reliability of that report should be increased more than in the case where the same number of sources agreed on the observation in the absence of supporting objects. This is because, in the latter case, it is harder for sources who propagate the content to verify it, unless they witnessed it themselves. This lends more doubt to the veracity of the propagated claims. Our new maximum likelihood estimator extends fact-finding approaches based on analysis of information propagation by automatically correlating the propagation with content features, allowing joint assessment of feature relevance and observation veracity, leading to improved estimation results.

Another significant direction in prior work uses training (from labeled data) to understand how content features correlate with veracity on social media [15], [16], [17]. They mainly adopt a *supervised* learning approach. For instance, Castillo *et al.* [15] adopted a decision tree model to detect fake news on Twitter based on various features of tweets. A disadvantage of supervised learning is that data should be labeled, costing a lot of human labor. In contrast, we develop an *unsupervised* method, where the importance of various features for truth discovery in social sensing is learned automatically with no need for labelling.

This paper is an extension of work originally presented in FUSION 2019 [18]. This work is different from the prior paper in several respects. First, we propose a new fact-finding algorithm, called constrained expectation maximum likelihood with multiple features (CEM-MultiF), which builds on the previous CEM algorithm to incorporate multiple corroborating features as a constraint to boost the probability of correctness of some claims. Second, we explore how the penalty factor, α, influences the estimation accuracy of the PEM-MultiF algorithm and show how to choose the best penalty factor to do experiments. To better evaluate the performance of the proposed algorithms, three extra metrics, recall, precision and F1, are adopted in this paper. We also conduct additional experiments with a new data set (US-Russia relation) to compare the performance of the proposed algorithms with baseline methods. We further discuss the advantages and disadvantages of different algorithms for truth discovery. We then review and summarize related work on truth discovery in recent years.

Finally, we discuss the possible presence of malicious sources that aim to subvert fact-finding by artificially adding veracity-correlated features to the content. In this context, we evaluate two proposed algorithms, called EM-MultiF and PEM-MultiF, demonstrating that both outperform the baselines for truth discovery in social sensing, and offer a performance/robustness trade-off in the presence of malicious sources. Also, we show that the PEM-MultiF algorithm has a higher estimation accuracy than the EM-MultiF method, especially when some true observations are reported by a small number of sources.

The rest of the paper is organized as follows. Section 2 reviews our model of truth discovery in social sensing applications. We develop two novel maximum likelihood algorithms that jointly assess feature relevance and content veracity in Section 3. In Section 4, we evaluate the estimation accuracy of the proposed algorithms using real-world data sets. Section 5 summarizes related work on truth discovery. We conclude the paper in Section 7.

2 Preliminaries

In this section, we review our model of the truth discovery problem in social sensing.

Consider an online social medium, such as Twitter, where sources report observations regarding the physical environment. Assume that a group of N sources (in some collected data stream), denoted $\mathcal{S} = \{S_1, S_2, \ldots, S_N\}$, report a total of M different assertions. For the purposes of this paper, we consider the information content of a tweet to be an assertion. Similar tweets make the same assertion. An assertion is thus a statement (in text), potentially supported by images, URLs, emoji, or other features, as will be described below. In general, multiple sources may report the same assertion. Let C_j denote the (unknown) Boolean variable that indicates whether the jth assertion is true or not. If the assertion is *true*, we say that $C_j = 1$. Otherwise, $C_j = 0$, if it is *false*. We overload the meaning of C_j to also denote the statement of the jth assertion, when no confusion arises. The act of reporting an assertion by a source is called a *claim* made by that source. When assertion C_j is reported by source S_i, we say that S_i claims C_j, or $S_iC_j = 1$. Otherwise, $S_iC_j = 0$, indicating that source S_i does not claim C_j. The elements S_iC_j can thus be stored into a two dimensional ($N \times M$) matrix, called the observation matrix, SC.

Different from prior work, this paper considers other features of tweets, such as the presence of images or URLs. Ideally, semantic analysis of such content can improve fact-finding. In this paper, however, we do not actually inspect the content of the image or URL referred to in a tweet. Rather, we consider binary indicators only, such as whether an image or URL was present. The reason is akin to crowdsourcing; individuals who propagate the tweet would have usually had a chance to see the corroborating content it refers to. The resulting propagation pattern then encodes what those individuals thought of the tweet when presented with its corroborating content. Their behavior gives the fact-finder additional information it can use to infer content veracity. Let K denote the number of features that we consider (that can be extracted from assertions). We denote the kth feature by F_k. When the kth feature in assertion C_j is present, we say that $F_kC_j = 1$. Otherwise, we say that $F_kC_j = 0$. Note that, in making assertion C_j, it may be that some sources include a feature, F_k, whereas others do not. For example, in tweeting about the same car accident, some sources may include a picture, whereas others not. We say that $F_kC_j = 1$ if the *majority* of sources (it means more than half of sources) reporting C_j include feature F_k. Hence, the elements, F_kC_j, could be collected in a two dimensional ($K \times M$) matrix. We call it the feature matrix, *FC*.

Finally, sources on social media may be connected by an influence graph. Prior work offers different ways to estimate

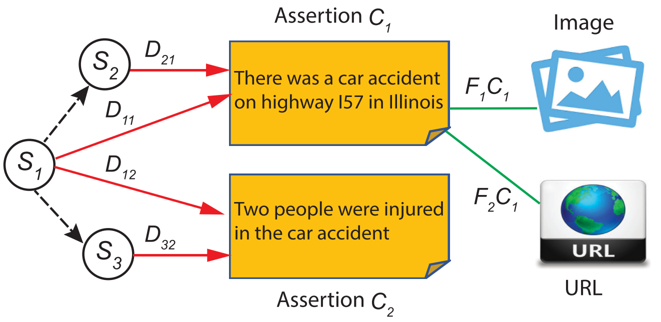


Fig. 1. Illustration of sources are connected by social graphs. S_1 first reports two assertions C_1 and C_2, then S_2 and S_3 follow S_1 to retweet. In addition, assertion C_1 is reported together with Image and URL.

the influence graph empirically from retweet patterns (in a nutshell, a source is considered to be influenced by another if they retweet them with a sufficient frequency). Fig. 1 shows an influence graph. A source S_i is called a successor of source S_t if there exists an edge from S_t to S_i in the graph. We call sources downstream from S_t its *descendants*, and call S_t their *ancestor*. Those with no incoming edges are root sources. We introduce the indicator D_{ij} to denote whether a certain source S_i has ancestors who report assertion C_j or not. Let $D_{ij} = 1$ denote that some ancestors of S_i make assertion C_j. $D_{ij} = 0$ represents that no ancestor of S_i makes assertion C_j. We further define a dependency matrix, D, as the two-dimensional matrix of all such indicators.

Note that, if $S_iC_j = 1$ and $D_{ij} = 1$, then source S_i might not be acting independently in claiming C_j (hence, the name *dependency* matrix). Rather, it may simply be repeating C_j because an ancestor of theirs made the same assertion (by definition of $D_{ij} = 1$). This has implications on veracity analysis. Specifically, when considering an assertion made by a source, we need to consider whether their ancestors made the same assertion. Fig. 1 shows indicators D_{ij} for three sources making two assertions. In this case, $D_{11} = 0$ because source S_1 does not have any ancestors who assert C_1 (which in this figure refers to the assertion: “There was a car accident on the highway I57 in Illinois”). On the other hand, $D_{21} = 1$ because source S_2 has an ancestor S_1 who makes assertion C_1. Also, $D_{32} = 1$ because source S_3 has an ancestor, S_1, who makes assertion C_2.

The final goal of veracity analysis is to estimate the unknown truth values C_j for each assertion. We do so jointly with estimating the different veracity indicators in an unsupervised manner (that does not require any data labeling or prior statistical training to estimate important of different features). Rather, inspired by the method in [14], we define the following indicators 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 ancestor 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.
- $p_k = P(F_kC_j = 1|S_iC_j, C_j = 1)$: The probability that assertion C_j includes feature F_k (i.e., that the majority of sources making that assertion would include that feature), given that assertion C_j is true.
- $q_k = P(F_kC_j = 1|S_iC_j, C_j = 0)$: The probability that assertion C_j includes feature F_k, given that assertion C_j is false.

It remains to show how to estimate the values of the above indicators together with the unknown truth values, C_j. This is done using a maximum likelihood estimation approach.

3 Expectation Maximization With Multiple Features

In this section, we first develop the new expectation maximization (EM) integrated with multiple features, called EM-MultiF, for truth discovery in Section 3.1. Then Section 3.2 proposes a novel penalized expectation maximization (PEM) integrated with multiple features, called PEM-MultiF, to improve the estimation accuracy of truth discovery.

3.1 Expectation Maximization With Multiple Features

For our problem of truth discovery, the latent variables to be determined are the set, $C = \{C_1, C_2, \ldots, C_M\}$, denoting whether the respective assertions are true or false. They will be determined based on the observation matrix, SC, the feature matrix, FC, and the dependency matrix, D. Let d denote the unknown expected fraction of correct assertions, $P(C_j = 1)$, in the data set.

Let us define θ to be the vector of all unknown parameters, $\theta = [d; a_i, b_i, f_i, g_i, p_k, q_k]$. The goal of this paper is to estimate the unknown parameters θ together with the truth value of each assertion, C_j, given the observed data, SC, FC, and the social graph, D. This paper adopts an expectation maximization algorithm which integrates the source claims, SC, with multi-dimensional features, FC, to solve the truth discovery problem.

The EM starts with defining a log likelihood function that expresses the (log of the) likelihood of received observations as a function of parameters and latent variables to be estimated. For our problem, the log likelihood function is given by

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

Given the log likelihood, EM defines two iterative steps, called the E-step and M-step, that converge to a maximum likelihood estimate.

The E-step is

$$\begin{aligned}Q(\theta|\theta^t) = &\sum_{C\in\{0,1\}} P(C|SC, FC; D, \theta^t)\\ &\times \ln\left(P(SC, FC|C; D, \theta)P(C; D, \theta)\right).\end{aligned} \tag{2}$$

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

Since FC and SC are dependent in the Bayesian graphical model, the above Eq. (2) can be expressed by the following equation based on conditional probability.

$$Q(\theta|\theta^t) = \sum_{C \in \{0,1\}} P(C|SC, FC; D, \theta^t) \times \ln\Big(P(FC|SC, C; D, \theta)P(SC|C; D, \theta)P(C; D, \theta)\Big). \tag{3}$$

Since there exists M assertions in set C, the above Eq. (2) could be rewritten as

$$Q(\theta|\theta^t) = \sum_{j=1}^{M} \sum_{C_j \in \{0,1\}} P(C_j|FC_j, SC_j; D, \theta^t)\Big\{ \ln P(C_j; D, \theta) + \ln\Big(P(FC_j|SC_j, C_j; D, \theta)P(SC_j|C_j; D, \theta)\Big)\Big\}, \tag{4}$$

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

$$Z(j,t) = P(C_j|SC_j, FC_j; D, \theta^t) = \frac{P(FC_j|SC_j, C_j; D, \theta^t)P(SC_j|C_j; D, \theta^t)P(C_j; D, \theta^t)}{\sum_{C_j \in \{0,1\}} P(FC_j|SC_j, C_j; D, \theta^t)P(SC_j|C_j; D, \theta^t)P(C_j; D, \theta^t)}. \tag{5}$$

For the above Eq. (5), SC_j refers to the jth claim made by N sources, so we have

$$P(SC_j|C_j; D, \theta^t) = \prod_{i=1}^{N} P(S_iC_j|C_j; D, \theta^t). \tag{6}$$

And FC_j refers to the jth claim reported by sources with K features, so we can have

$$P(FC_j|SC_j, C_j; D, \theta^t) = \prod_{k=1}^{K} P(F_kC_j|SC_j, C_j; D, \theta^t), \tag{7}$$

where $P(S_iC_j|C_j; D, \theta^t)$ and $P(F_kC_j|SC_j, C_j; \theta^t)$ are corresponding to the parameters $a_i, b_i, f_i, g_i, p_k, q_k$ defined in Section 2 given different D and C_j.

Substituting Eqs. (6) and (7) into Eq. (4), yielding

$$Q(\theta|\theta^t) = \sum_{j=1}^{M} \sum_{C_j \in \{0,1\}} P(C_j|FC_j, SC_j; D, \theta^t)\Big\{ \ln P(C_j; D, \theta) + \sum_{i=1}^{N} \ln P(S_iC_j|C_j; D, \theta) + \sum_{k=1}^{K} \ln P(F_kC_j|SC_j, C_j; D, \theta)\Big\}. \tag{8}$$

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, yielding

$$\theta^{t+1} = \arg\max Q(\theta|\theta^t). \tag{9}$$

To solve the above Eq. (9), we can take the gradient of the parameters θ in (8) and then make them equal to 0. So we have

$$\frac{\partial Q(\theta|\theta^t)}{\partial a_i} = \frac{\sum_{C_j \in S_iC_1^{D_0}} Z_j}{a_i} - \frac{\sum_{C_j \in S_iC_0^{D_0}} Z_j}{1 - a_i}, \tag{10a}$$

$$\frac{\partial Q(\theta|\theta^t)}{\partial b_i} = \frac{\sum_{C_j \in S_iC_1^{D_0}} (1 - Z_j)}{b_i} - \frac{\sum_{C_j \in S_iC_0^{D_0}} (1 - Z_j)}{1 - b_i}, \tag{10b}$$

$$\frac{\partial Q(\theta|\theta^t)}{\partial f_i} = \frac{\sum_{C_j \in S_iC_1^{D_1}} Z_j}{f_i} - \frac{\sum_{C_j \in S_iC_0^{D_1}} Z_j}{1 - f_i}, \tag{10c}$$

$$\frac{\partial Q(\theta|\theta^t)}{\partial g_i} = \frac{\sum_{C_j \in S_iC_1^{D_1}} (1 - Z_j)}{g_i} - \frac{\sum_{C_j \in S_iC_0^{D_1}} (1 - Z_j)}{1 - g_i}, \tag{10d}$$

$$\frac{\partial Q(\theta|\theta^t)}{\partial p_k} = \frac{\sum_{C_j \in F_kC_1} Z_j}{p_k} - \frac{\sum_{C_j \in F_kC_0} Z_j}{1 - p_k}, \tag{10e}$$

$$\frac{\partial Q(\theta|\theta^t)}{\partial q_k} = \frac{\sum_{C_j \in F_kC_1} (1 - Z_j)}{q_k} - \frac{\sum_{C_j \in F_kC_0} (1 - Z_j)}{1 - q_k}, \tag{10f}$$

$$\frac{\partial Q(\theta|\theta^t)}{\partial d} = \frac{\sum_{i=1}^{M} Z_j}{d} - \frac{\sum_{i=1}^{M} (1 - Z_j)}{1 - d}. \tag{10g}$$

where $Z_j = P(C_j = 1|SC_j, FC_j; D, \theta^t)$, $S_iC_1^{D_0} = \{S_iC_j : \forall S_iC_j \ \&S_iC_j = 1 \ \&D_{ij} = 0\}$, $S_iC_0^{D_0} = \{S_iC_j : \forall S_iC_j \in SC \ \&S_iC_j = 0 \ \&D_{ij} = 0\}$, $S_iC_1^{D_1} = \{S_iC_j : \forall S_iC_j \in SC \ \&S_iC_j = 1 \ \&D_{ij} = 1\}$, $S_iC_0^{D_1} = \{S_iC_j : \forall S_iC_j \in SC \ \&S_iC_j = 0 \ \&D_{ij} = 1\}$, $F_kC_1 = \{F_kC_j : \forall F_kC_j \in FC \ \&F_kC_j = 1\}$, $F_kC_0 = \{F_kC_j : \forall F_kC_j \in FC \ \&F_kC_j = 0\}$. M is the total number of assertions reported by sources.

Let the gradient of each parameter in Eq. (10) be 0 and then we obtain the answers below

$$a_i^{t+1} = \frac{\sum_{C_j \in S_iC_1^{D_0}} Z_j}{\sum_{C_j \in S_iC_1^{D_0} \cup S_iC_0^{D_0}} Z_j}, \tag{11a}$$

$$b_i^{t+1} = \frac{\sum_{C_j \in S_iC_1^{D_0}} (1 - Z_j)}{\sum_{C_j \in S_iC_1^{D_0} \cup S_iC_0^{D_0}} (1 - Z_j)}, \tag{11b}$$

$$f_i^{t+1} = \frac{\sum_{C_j \in S_iC_1^{D_1}} Z_j}{\sum_{C_j \in S_iC_1^{D_1} \cup S_iC_0^{D_1}} Z_j}, \tag{11c}$$

$$g_i^{t+1} = \frac{\sum_{C_j \in S_iC_1^{D_1}} (1 - Z_j)}{\sum_{C_j \in S_iC_1^{D_1} \cup S_iC_0^{D_1}} (1 - Z_j)}, \tag{11d}$$

$$p_k^{t+1} = \frac{\sum_{C_j \in F_kC_1} Z_j}{\sum_{C_j \in F_kC_1 \cup F_kC_0} Z_j}, \tag{11e}$$

$$q_k^{t+1} = \frac{\sum_{C_j \in F_kC_1} (1 - Z_j)}{\sum_{C_j \in F_kC_1 \cup F_kC_0} (1 - Z_j)}, \tag{11f}$$

$$d^{t+1} = \frac{\sum_{j=1}^{M} Z_j}{M}. \tag{11g}$$

An estimate of the unknown parameters θ can be thus obtained from Eq. (11) above. The E-steps and M-steps can be solved iteratively until they converge. A pseudo-code is presented later in the paper.

3.2 Penalized Expectation Maximization

For the former EM-MultiF algorithm, we assume the different features are conditionally independent. In reality, they are not [19]. Empirical data suggests that the probability that an assertion is false when multiple truth indicators are present (e.g., a supporting URL and a supporting image) is a lot smaller than what would be expected if these indicator features were independent. As a result, the EM-MultiF algorithm may not lead to the best solution. To deal with this issue, we present another novel algorithm, penalized expectation maximization, integrated with multiple features (PEM-MultiF), to improve the estimation accuracy for truth discovery in social sensing.

Different from the former EM-MultiF method, we penalize the posterior probability of latent variables when an assertion includes multiple corroborating features, such as images and URLs. As mentioned above, an assertion is more likely to be true when it is supported by multiple features, compared to what EM-MultiF computes based on its implicit feature independence assumption. Therefore, we add a penalty to discount the posterior probability that an assertion is false when it is supported by more than one corroborating feature. The penalty increases with the number of seen features.

For each assertion j, sources will report it with K features, so the above maximum likelihood function in Eq. (1) could be rewritten as

$$\begin{aligned}\mathcal{L}_p &= \ln P(SC, FC; D, \theta)\\ &= \ln\Bigg(\sum_{C_j\in\{0,1\}}\prod_{j=1}^{M}\prod_{k=1}^{K} P(F_kC_j|SC_j, C_j; D, \theta)\\ &\quad \times P(SC_j|C_j; D,\theta)P(C_j; D,\theta)\Bigg).\end{aligned} \tag{12}$$

As we mentioned above, when an assertion is reported by sources with multiple features, its probability to be false is pretty small. In addition, the different features are possibly correlated with each other. Thus, we could penalize the term $\prod_{k=1}^{K} P(F_kC_j|SC_j, C_j; D, \theta)$ with respect to $C_j = 0$ while keeping it unchanged as $C_j = 1$. Let α $(0 < \alpha < 1)$ denote the penalty factor and n_j be the number of features that the jth assertion has, $n_j = \|FC_j\|_1$. Then Eq. (12) could be reformulated as

$$\begin{aligned}\mathcal{L}_p = \ln\Bigg(&\sum_{C_j\in\{0,1\}}\prod_{j=1}^{M}\prod_{k=1}^{K} \alpha^{n_j(1-C_j)} P(F_kC_j|SC_j, C_j; D, \theta)\\ &\times P(SC_j|C_j; D,\theta)P(C_j; D,\theta)\Bigg).\end{aligned} \tag{13}$$

For the above Eq. (13), when $C_j = 1$, the penalized term $\alpha^{n_j(1-C_j)}$ becomes 1, indicating no penalty. However, it will be penalized when $C_j = 0$ if assertion j is reported by sources with multiple features together. In addition, when $n_j = 0$, it indicates that there are no other features supporting assertion j made by sources. This assertion can be either true or false, depending on the reliability of sources. When $n_j \geq 1$, the penalized term works to lower the probability of the assertion to be false, because the assertion is more likely to be true when it is supported by multiple features. Note that the larger n_j, the lower the probability of sources making error.

Next, we derive the E-step of the above penalized likelihood function as follows:

$$\begin{aligned}Q(\theta|\theta^t) = \sum_{j=1}^{M}\sum_{C_j\in\{0,1\}} Z(j,t)\Bigg\{& n_j(1-C_j)\ln\alpha + \ln P(C_j; D, \theta)\\ &+\sum_{k=1}^{K}\ln P(F_kC_j|SC_j, C_j; D,\theta) + \sum_{i=1}^{N}\ln P(S_iC_j|C_j; D,\theta)\Bigg\},\end{aligned} \tag{14}$$

where $Z(j,t)$ is the posterior probability of the latent variable C_j, which is given by

$$\begin{aligned}&Z(j,t) = P(C_j|SC_j, FC_j; D, \theta^t) =\\ &\frac{\alpha^{n_j(1-C_j)}P(FC_j|SC_j, C_j; D, \theta^t)P(SC_j|C_j; D,\theta^t)P(C_j)}{\sum_{C_j\in\{0,1\}}\alpha^{n_j(1-C_j)}P(FC_j|SC_j, C_j; D, \theta^t)P(SC_j|C_j; D,\theta^t)P(C_j)},\end{aligned} \tag{15}$$

where $P(SC_j|C_j; D, \theta^t)$ and $P(FC_j|SC_j, C_j; D, \theta^t)$ are defined in Eqs. (6) and (7). They are corresponding to the parameters $a_i, b_i, f_i, g_i, p_k, q_k$ defined in Section 2.

Then, the M-step is used to maximize the $Q(\theta|\theta^t)$ to estimate the unknown parameter θ given the posterior of latent variable. It can be expressed by

$$\theta^{t+1} = \arg\max Q(\theta|\theta^t). \tag{16}$$

Using the same methodology in Section 3.1, the M-step leads to the same updates as Eq. (11) except that $Z(j,t)$ is from Eq. (15) rather than Eq. (5). Note that, here we do not present the detailed equations like Eq. (11) in order to avoid redundancy.

3.3 Constrained Expectation Maximization

In this subsection, we introduce a new constrained expectation maximization algorithm that uses multiple content features (CEM-MultiF) to boost accuracy of truth discovery. This algorithm builds on the prior constrained maximum likelihood estimator (CEM) [14]. However, the CEM algorithm only takes users' posting behaviors into account in truth estimation, without considering content features. In contrast, in this paper, we incorporate information on different content features to constrain the probability distributions of latent variables. Our mathematical insight is that a claim is more likely to be true if it is supported by multiple independent corroborating features. Assume that a user made a claim supported by corroborating evidence. Let λ denote the probability of a chance agreement with that a claim by another user who independently cites a different type of corroborating evidence. The probability that the original claim is true when supported by N such subsequent independent pieces of evidence is thus $1-\lambda^N$ (i.e., the probability that at least one of the pairwise agreements was not coincidental). Observe that the total number of corroborating pieces here is $N+1 = n_j$. Hence, we prevent the expectation maximization algorithm from dipping below that above probability by adding the constraint

$$1-\lambda^{n_j-1} \leq q(C_j|SC_j) \leq 1, \tag{17}$$

where n_j is the number of independent corroborating features. Thus, the probability of correctness of claims in this model will depend in part on content features. Using a similar methodology to CEM, the E-step in the CEM-MultiF algorithm is defined by

$$\begin{aligned} &\underset{q(C_j|SC_j)}{\arg\min}\ \ KL\big(q(C_j|SC_j)||P(C_j|SC_j; D, \theta^t)\big), \\ &\text{s.t.,}\ \forall_j :\ \ 1 - \lambda^{n_j - 1} \leq q(C_j|SC_j) \leq 1, \end{aligned} \tag{18}$$

where $q(C_j|SC_j)$ is the posterior probability of correctness of claims, n_j is the number of content features from claims and λ is the hyper-parameter.

For the above formula, we first calculate the posterior probability, $q(C_j|SC_j)$, based on the same methodology as Eq. (5). When a claim is made with multiple features and $q(C_j|SC_j)$ is smaller than the constraint probability, $1 - \lambda^{n_j - 1}$, we constrain the probability of correctness as $1 - \lambda^{n_j - 1}$. Note that, if $q(C_j|SC_j)$ is larger than the constraint probability, we do not need to update it. The next step is to use the same M-step as EM-MultiF to maximize the unknown parameters given the posterior probability above. Finally, we repeat the E-step and M-step alternatively until they converge.

Algorithm 1. EM, CEM/PEM-MultiF for Truth Discovery

```
1:  Input: initialize $\theta$ with reporting ratio, $n_j$, $d = 0.5$
2:  Output: Classification results: $\hat{C}_j$, $\theta$
3:  while $\theta^t$ does not converge do
4:    for $j = 1 : M$ do
5:      Compute posterior $Z(j, t)$ based on Eqs. (5), (15) and (18)
6:      for EM-MultiF, PEM-MultiF and CEM-MultiF,
        respectively
7:    end for
8:    for $i = 1 : N$ do
9:      Compute $a_i^{t+1}, b_i^{t+1}, f_i^{t+1}, g_i^{t+1}, p_k^{t+1}, q_k^{t+1}, d^{t+1}$ based on
        Eq. (11)
10:   end for
11:   Update $\theta^{t+1} = \theta^t$
12:   $t = t + 1$
13: end while
14: for $j = 1 : M$ do
15:   if $Z(j, t) \geq 0.5$ then
16:     $\hat{C}_j$ = 1 (true)
17:   else
18:     $\hat{C}_j$ = 0 (false)
19: end for
20: Return classification results $\hat{C}_j$
```

3.4 Algorithm

In this subsection, we summarize the two novel maximum likelihood algorithms, EM-MultiF and PEM-MultiF, for truth discovery problem above, as shown in Algorithm 1.

In Algorithm 1, parameters in the set θ except for d are first initialized with the reporting rate, which is defined as the ratio of the number of assertions (content features) reported by ith user to the total assertions reported by all the users. The unknown expected fraction of correct assertions, d, is initialized to 0.5. After initializing, we compute the posteriors of latent variables, C_j, through the E-step and then calculate the unknown parameters θ in the M-step using the maximum likelihood algorithm. Lines 5 and 6 compute the posterior probability of the latent variables for EM-MultiF and PEM-MultiF algorithm, respectively. After the parameters θ converge, we classify the assertions based on the estimated results. If the calculated probability of an assertion is equal or greater than 0.5, the assertion is said to be true, $\hat{C}_j = 1$; otherwise it is said to be false, $\hat{C}_j = 0$. Note that, the difference between EM-MultiF and PEM-MultiF is in how the posterior probability of latent variables is computed. In PEM-MultiF, we penalize the posterior of an assertion when it is supported by multiple features.

4 Performance Evaluation

In this section, we evaluate the performance of the proposed EM-MultiF and PEM-MultiF algorithms for truth discovery on the real-world data sets collected from Twitter. We first describe the data sets used, our ground-truthing methods, as well as data preprocessing used to pre-compute the inputs to our algorithm. We also explain the compared baselines. We then present three sets of experimental evaluation results to demonstrate the effectiveness of the proposed algorithms. Finally, we explore limiting the number of penalty times for each source to impede malicious sources.

4.1 Data Collection and Groundtruthing

We collected real-world data sets from Twitter through our developed Twitter crawler system, called *Apollo*,[1] that allows users to collect tweets by typing keywords. For example, if we input the keyword "hurricane" with geo-location, it can capture all the tweets about hurricanes at that location. In this study, we collect three real-world data traces related to *Hurricane Harvey in Texas*, *US-Russia News* in 2017 and *Russia-Ukraine* conflict in 2019 from Twitter. Table 1 briefly summarizes statistics of these three data traces. More background is presented below:

- Hurricane Harvey: On Aug. 25th 2017, Hurricane Harvey made landfall in Texas. It became the country's first Category 4 storm since Wilma hit Florida in October 2005. This tropical storm caused about $125 billion in damage.
- US-Russia: Tweets about the relation between the United States and Russia in Aug. 2017. There were allegations of Russian manipulation spreading over social media.
- Ukraine: Tweets related to the Ukraine-Russia conflict, collected from Jan. 20 14:33:19, 2019 to Feb. 08 07:52:21, 2019.

We use the most popular 600 assertions from each data set to evaluate the performance of the proposed algorithms. These assertions are reported by many different users on Twitter. We did not consider the remaining assertions because they are only reported/propagated by very few sources, and thus are not perceived as important. We further use only those sources that make at least two tweets during the evaluation period. This is because it is harder to reason about veracity of inactive sources. Finally, we extract three important features from the tweets: images, URLs and

[1] Online. [Available]: http://apollo4.cs.illinois.edu/

TABLE 1
Information Summary About Three Data Sets Collected From Twitter

Data set	Start Time (UTC)	End Time (UTC)	# Tweets	# Assertions	# URLs	# Images	# Sources
Hurricane Harvey	Aug 26 11:05:26, 2017	Aug 27 22:23:31, 2017	59,620	15,946	10,706	2,522	50,054
US-Russia Relation	Aug 02 23:30:04 2017	Aug 04 10:09:02 2017	65,120	21,716	15,581	2,858	46,168
Ukraine	Jan 20 14:33:19, 2019	Feb 08 07:52:21, 2019	72,963	10,206	9,782	240	39,501

claims reported by at least two independent sources, to feed our model with binary measurements. For example, if an assertion is reported by at least two independent sources, the corresponding value in the feature matrix, FC, is set to 1.

In our experiments, human graders are asked to label the ground truth of each assertion (tweet) without knowing which algorithm generates the results in order to prevent bias. Graders mark the assertions as "True" and "False" according to the following rules:

- *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 and subjective comments made by sources.

To reduce the subjectivity of human graders, multiple graders are asked to mark the same assertions by searching several diverse media sources. Majority voting is then adopted to assess the final labels of the assertions. In case of a tie, we ask other graders to mark the assertion to break the tie.

4.2 Data Preprocessing

In this subsection, we describe how to preprocess the collected data sets from our developed Twitter crawler system, *Apollo*. The raw data download using the native Twitter API contains many features such as user IDs, tweets and retweets, URLs, image pointers, and the number of followers. We first adopt cosine similarity [20] to cluster similar tweets into assertions. Specifically, we first remove some special and irrelevant characters from the post content. We then use the cosine similarity function in the Python package "scikit-learn"[2] to cluster the similar claims. For the follower-followee relationship, if source S_i who follows S_j reports the same assertion at least two times, they would be deemed to have a dependency relationship, $D_{ij} = 1$; otherwise, $D_{ij} = 0$. In addition, we further extract the binary values of URLs and images from tweets, and whether the claims reported by at least two independent sources to construct feature claim (FC) matrix. In the FC matrix, we use binary measurements to denote whether a feature, F_k, in the jth assertion, C_j, is reported by the majority sources.

4.3 Baseline Methods

To evaluate the performance of the proposed PEM-MultiF and EM-MultiF algorithms in this paper, we compare them with the existing fact-finding methods below.

- *CEM-MultiF*: This is a newly proposed baseline method in this paper that builds on the CEM algorithm in INFOCOM'18 [14]. It replaces the number of independent sources in the constraint with multi-dimensional features to boost estimation performance.
- *CEM*: This algorithm was proposed in INFOCOM'18 [14]. It incorporates prior information about the number of independent sources into the model to constrain the probability of latent truth variables.
- *EM-social*: This algorithm was proposed in IPSN'14 [4]. It uses the general EM algorithm to estimate the correctness of assertions given the social graph.
- *EM-regular*: This algorithm was proposed in IPSN'12 [21] and uses the general EM algorithm to evaluate the truth values of assertions, *without* considering the social graph.

4.4 Effect of α and λ on PEM-MultiF and CEM-MultiF

Effect of Penalty Factor α on PEM-MultiF . In our first experiment, we tune the penalty factor, α, and explore how it influences the estimation accuracy of the PEM-MultiF algorithm using the above real-world data sets. This paper uses *US-Russia* data set as the training data and the other two data traces as the testing data. Specifically, we first use the *US-Russia* data set to tune the best penalty factor, α, and then apply it to the testing data to check the result. Fig. 2 shows the estimation accuracy of the PEM-MultiF algorithm under different penalty factor, α. We can observe that it has the highest estimation accuracy when $\alpha = 0.01$ for the training data set, *US-Russia* data. When $\alpha = 0.01$, its estimation accuracy for *US-Russia* data trace is about 80.5 percent, but it drops to 75 percent when $\alpha = 0.5$. Then we further adopt the penalty factor, $\alpha = 0.01$, to conduct experiments on the other two testing data sets, Hurricane Harvey and Ukraine. The results illustrate that the PEM-MultiF algorithm can achieve best performance when $\alpha = 0.01$. Therefore, we set penalty factor, α, to 0.01 as default value in the following experiments.

Effect of Constraint Factor λ on CEM-MultiF. Similarly, we explore how hyper-parameter, λ, influences the estimation

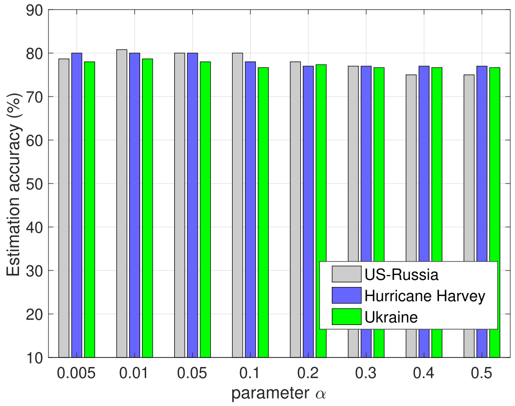


Fig. 2. Influence of α on the estimation accuracy of PEM-MultiF.

[2] Online. [Available]: https://scikit-learn.org/stable/

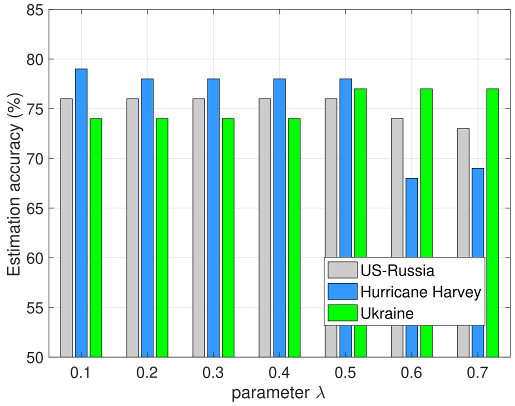


Fig. 3. Influence of λ on the estimation accuracy of CEM-MultiF.

accuracy of the CEM-MultiF algorithm. Fig. 3 illustrates the estimation accuracy of the CEM-MultiF algorithm under different parameter, λ. From this figure, it can be seen that CEM-MultiF performs best when $\lambda = 0.4$ and 0.5 for *US-Russia* data set. Then we further verify these two values on other two testing data sets. We can still find that CEM-MultiF achieves very high accuracy as $\lambda = 0.5$. Therefore, we choose $\lambda = 0.5$ as default value to conduct the following experiments.

4.5 Empirical Results

Next, we conduct extensive experiments to compare the proposed algorithms with the existing fact-finding algorithms based on real data sets above. This work adopts four widely used information retrieval metrics: accuracy, precision, recall, and F1 score. Accuracy refers to the ratio of correctly predicted assertions to the total assertions. Precision is defined as the percentage of assertions labeled "true" that were indeed true. Recall is the percentage of true assertions that are classified as such. Finally, F1 score takes both the recall and precision into account, defined as

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{19}$$

For the first experiment, we use the real-world data set, *Hurricane Harvey*. Fig. 4 illustrates the evaluation results for the most popular 100 labeled assertions, because users are mainly interested in hot topics. We can observe that the PEM-MultiF and EM-MultiF algorithms outperform the baselines in terms of estimation accuracy. The main reason is that we integrate source claims with multiple features into our model, which can better assess the correctness of the observations in social media. In addition, the PEM-MultiF algorithm has a higher estimation accuracy than the EM-MultiF algorithm. This is because PEM-MultiF penalizes the probability of the latent variable if an assertion is reported with multiple features together. In addition, CEM-MultiF performs better than the CEM algorithm in this data set. The main reason is that CEM-MultiF adopts multiple features to boost the estimation quality to find more true assertions reported by one independent source. Finally, we further evaluate the estimation quality of each algorithm using three more evaluation metrics: *recall*, *precision* and *F1* score as shown in Table 2. We observe that the PEM-MultiF algorithm has the highest F1 score while its precision and recall are comparable to the other algorithms. Also, the CEM-MultiF and EM-MultiF algorithms have higher precision and F1 score compared to the other baselines except that the recall of EM-MultiF is a little lower than the CEM method. We thus conclude that the proposed PEM-MultiF algorithm and EM-MultiF algorithm have better estimation quality than the baselines.

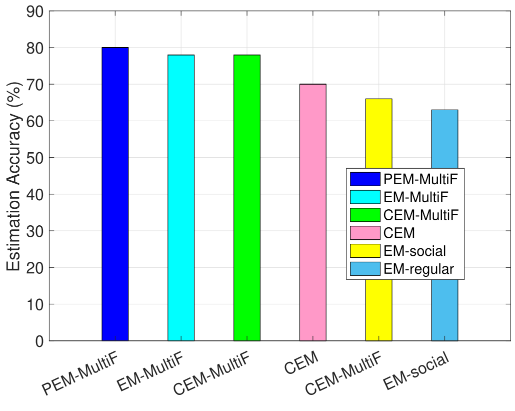


Fig. 4. Comparison of estimation accuracy for different algorithms (Harvey).

TABLE 2
Performance Evaluation Using Recall, Precision, and F1 (Harvey)

Algorithms	Recall	Precision	F1 score
PEM-MultiF	0.8481	0.89333	**0.87013**
EM-MultiF	0.74684	**0.96721**	0.84286
CEM-MultiF	**0.87342**	0.85185	0.8625
CEM	0.76136	0.8481	0.8024
EM-social	0.67089	0.86885	0.75714
EM-regular	0.64935	0.8928	0.75187

In addition, we evaluate the estimation accuracy of different algorithms on the real-world data set, *US-Russia* . Fig. 5 illustrates the accuracy comparison for different algorithms above. It can be seen that the PEM-MultiF algorithm still has the highest estimation accuracy compared to the other algorithms. The EM-MultiF and CEM-MultiF algorithms beat the other baselines in terms of estimation accuracy. Similarly, we further evaluate the performance of the

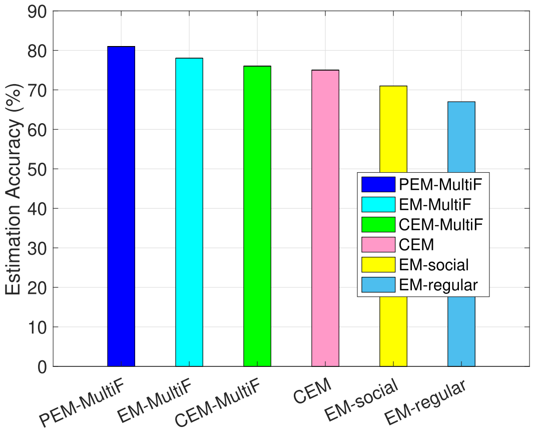


Fig. 5. Comparison of estimation accuracy for different algorithms (US-Russia).

TABLE 3
Performance Evaluation Using Recall, Precision, and F1 (US-Russia)

Algorithms	Recall	Precision	F1 score
PEM-MultiF	0.97368	0.81319	**0.88623**
EM-MultiF	0.86842	**0.84615**	0.85714
CEM-MultiF	**0.98592**	0.75269	0.85366
CEM	0.93421	0.78022	0.85030
EM-social	0.84211	0.79012	0.81529
EM-regular	0.8000	0.76923	0.78431

TABLE 4
Performance Evaluation Using Recall, Precision, and F1 (Ukraine Data Set)

Algorithms	Recall	Precision	F1 score
PEM-MultiF	0.93204	0.79339	**0.85714**
EM-MultiF	0.8835	**0.79825**	0.83871
CEM-MultiF	**0.96386**	0.7619	0.85106
CEM	0.95146	0.74809	0.83761
EM-social	0.9375	0.73529	0.82418
EM-regular	0.72973	0.78641	0.75701

above algorithms using three more metrics: recall, precision and F1 score, as shown in Table 3. We can observe from Table 3 that the recall, precision and F1 score of the PEM-MultiF and EM-MultiF algorithms are higher than the baselines, except that the recall of EM-MultiF is a little smaller than the CEM-MultiF.

Finally, we conduct another experiment on the real-world data set, *Ukraine* , with a time window of two weeks. Fig. 6 shows the comparison results of different algorithms in terms of estimation accuracy. We can observe that both PEM-MultiF and EM-MultiF still have a better estimation quality than the other methods. In addition, the estimation accuracy of the CEM-MultiF, CEM and EM-social are comparable because most assertions come from one original ancestor and do not have URLs. In other words, few constraints could be used to boost the performance of the CEM algorithm. As above, Table 4 illustrates the comparison results of different algorithms under three extra metrics: recall, precision and F1. We can observe that PEM-MultiF has higher F1 than the other methods while its recall and precision are close to the EM-MultiF and CEM. For EM-MultiF algorithm, it has higher precision and F1 than the existing baselines, except that its recall is smaller than the CEM.

In summary, based on the evaluation results of the real-world data traces above, we can conclude that the proposed PEM-MultiF and EM-MultiF can achieve better performance than the state-of-the-art baselines.

4.6 Robustness to Malicious Sources

In the above experiments, we have not considered the impact of malicious sources who may purposely fool our algorithms by exploiting various features mentioned above. Now, we artificially add malicious sources to report 5, 10 and 15 percent false assertions of 150 assertions in the real-world *Harvey* data set. Specifically, we choose a few unreliable sources as malicious sources to attack the proposed models. Each malicious source chooses 5 random false assertions to report with multiple artificially added features mentioned above. Fig. 7 shows the estimation accuracy for different algorithms under a varying ratio of malicious sources. We can observe that the estimation accuracy of PEM-MultiF, EM-MultiF and CEM algorithms gradually drops as the ratio of malicious sources increases from 0 to 15 percent. In addition, the estimation accuracy of the PEM-MultiF and CEM algorithms decreases faster than that of the EM-MultiF algorithm with the increase of malicious sources, which means that the EM-MultiF algorithm is more robust. Also, we find that EM-social and EM-regular remain almost unchanged with the varying ratio of malicious sources. This is because EM-social and EM-regular do not consider content features in their models. Thus, they are not affected by content manipulation. Nevertheless, PEM-MultiF and EM-MultiF still outperform the others with 15 percent malicious agents.

Next, we explore how to lower the impact of malicious sources using the *Harvey* data set with 10 percent false assertions reported by malicious sources. We change the limit on the number of times that α^{n_j} penalty term can be used for each source, from 1 to ∞. We apply the penalty to claims with the largest number of sources first until the limit is reached per a source. The results are shown in Fig. 8. We can observe that the estimation accuracy of the PEM-MultiF and EM-MultiF algorithm is higher than the other algorithms when the $limit = 1$. In addition, their estimation accuracy gradually decreases when increasing the limit on

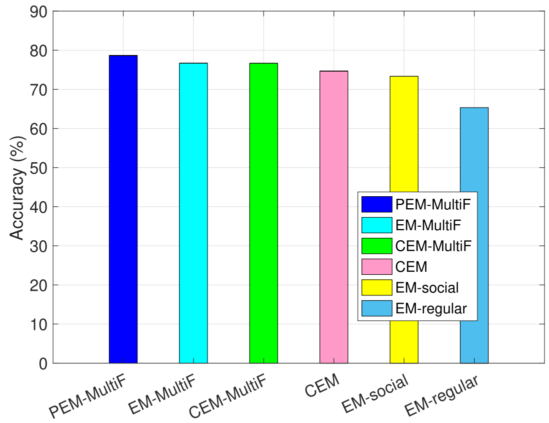


Fig. 6. Comparison of estimation accuracy for different algorithms (Ukraine data set).

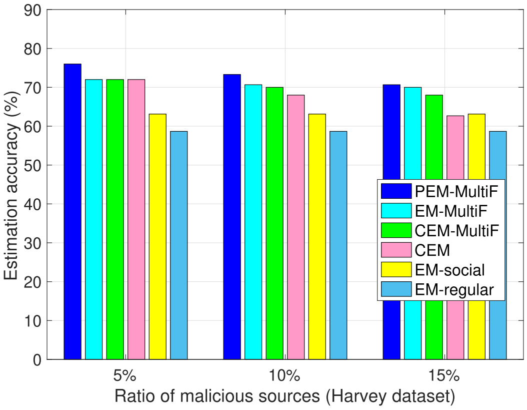


Fig. 7. Comparison of accuracy for different algorithms with malicious sources (Harvey data set).

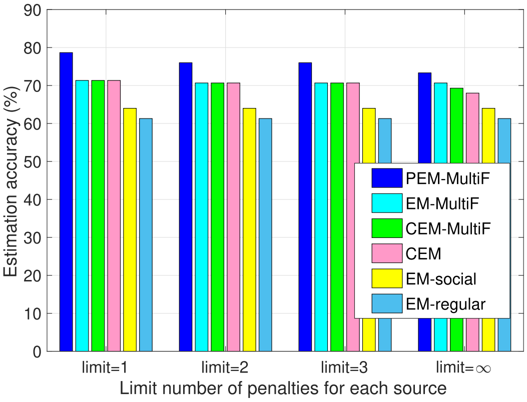


Fig. 8. Limit use of penalties for each source (10 percent malicious sources).

penalties. The main reason is that the higher the limit, the more opportunities malicious sources have to fool our algorithms. Therefore, limiting the use of penalty for each source to at most one time is best for our data set.

4.7 Discussion of Different Algorithms

Finally, we discuss the advantages and disadvantages of different algorithms above. Based on the above evaluation, we can see that PEM-MultiF has the highest estimation accuracy but it is vulnerable to malicious sources. The CEM-MultiF and EM-MultiF algorithms have higher estimation accuracy than the other baselines when assertions are reported (that include multiple corroborating features) by many sources. The EM-MultiF algorithm is more robust to malicious sources compared to the PEM-MultiF and CEM-MultiF algorithms. PEM-MultiF, CEM-MultiF and CEM algorithms tend to estimate more claims as being true, which has the effect of trading off precision for recall. CEM-MultiF has the higher recall but worse precision than PEM-MultiF due to the fact that it is very likely to perceive the claims to be true if they include multiple corroborating features. In addition, CEM-MultiF algorithm performs better than the CEM algorithm when true assertions are reported by one original source but include multiple corroborating features. This is because CEM-MultiF can boost the probability of correctness of such claims.

5 Related Work

In the past few years, truth discovery has attracted a lot of reseachers' attention. Yin *et al.* [22] introduced a unsupervised method, called *TruthFinder*, to iteratively estimate the true values of conflicting data from different sources on websites. Then several related approaches [23], [24], [25], [26] were developed to enhance the basic framework. For truth finding on social media, Wang *et al.* [21] first proposed an expectation maximization algorithm to jointly estimate the reliability of sources and evaluate the correctness of observations based on binary measurements. Wang *et al.* [4] later introduced a source depency model that improved the estimation of reliability of sources and the veracity of assertions by accounting for correlated errors (i.e., rumors) that spread along source dependency chains. Further, Shao *et al.* [14] developed a constrained expectation maximization (CEM) algorithm to constrain the posterior probability of an assertion reported by multiple independent sources to boost the likelihood of correctness.

Other researches derived error bounds on reliability. Xiao *et al.* [27] 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. Thus, some works [28] take source dependency into account for truth discovery. Ma *et al.* [29] proposed a iterative Expectation Maximization algorithm for Truth Discovery, called IEMTD, that jointly refers the reliability of agents and truth of events with dependent agents.

To better improve the data reliability, some work addressed the source selection problem. For instance, Shao *et al.* [30] formulated a non-linear integer programming problem to select optimal sources to solicit in order to minimize the expected fusion error. Amintoosi *et al.* [31] developed a privacy-aware participant selection framework to protect users' privacy in social networks. They considered the situation in which sources are incentivized to do tasks.

Some researchers also used supervised learning based on content features for truth discovery in social networks [32], [33], [34]. For instance, Gupta *et al.* [35] proposed a decision tree classifer to indentify fake images on Twitter effectively. Castillo *et al.* [15] assessed the credibility of news propagated through Twitter by using many features of tweets. In addition, some studies tend to use semi-supervised graph neural networks [36], [37], [38], [39] which encodes both the graph structure and node features to improve the detection accuracy. However, the drawback of supervised learning is that a lot of data should be labelled, costing lots of human labors.

This paper develops an unsupervised truth-finding approach with multi-modal data to improve the estimation accuracy of the existing methods in social sensing. Further, a heuristic penalty is introduced to boost the posterior probability of an assertion reported with multiple features together. We show that the proposed algorithms can significantly improve the ground truth estimation accuracy in social sensing.

6 Discussion and Future Work

In this section, we discuss limitations of presented work and outline opportunities for improvement. One potential limitation of this work, as presented, is the rather minimalist approach taken to the description of corroborating content features. Only binary indicators are considered that indicate whether images or URLs are present, but do not hint at properties of these images and URLs. More descriptive indicators that include some semantic description of corroborating features may lead to further improvements.

The constraints introduced (beyond the original maximum likelihood estimation) on the probability of correctness of latent variables are heuristic in nature. While intuitions were presented to back up these heuristics, it may be interesting to offer better analytical foundations to derive more accurate forms of these constraints.

Another limitation of our algorithms is that it is hard to apply them to veracity analysis of small events and claims. Instead, the approach works best when analyzing widely-disseminated claims. This, perhaps, is not a big problem because widely-disseminated claims are, by definition, the ones that garner more attention. Ascertaining veracity of such claims may be more important because the potential for damage from a piece of undetected misinformation is higher when this piece is more popular. This is precisely where our algorithms add more value.

Finally, we do not do any semantic analysis of text. Clearly, it is possible for the same meaning to be expressed quite differently by different sources (e.g., "The building is on fire" versus "The house is burning"). Our current cosine similarity metric will fail to identify such semantic similarity, as it compares claims only lexically. As a result, the two sentences in the above example will be considered different. Failure to recognize that the two, in fact, support each other will bias fact-finding towards underestimating veracity. Fortunately, the cosine similarity module is an isolated plug-and-play component of our architecture. It can be easily replaced by more advanced techniques that rely on word embedding or other methods from statistical linguistics to better assess similarity. This remains a topic for future work.

7 Conclusion

This paper developed novel expectation maximization algorithms integrated with multiple features, called PEM-MultiF, EM-MultiF and CEM-MultiF, to improve the accuracy of truth discovery in social sensing. For EM-MultiF, it incorporated multi-dimensional data from social media into the maximum likelihood model to estimate the correctness of the observations. The PEM-MultiF algorithm penalized the probability of an assertion to be false when it is supported by multiple features. For CEM-MultiF, it incorporated the information on multiple corroborating features as a constraint to boost the posterior probability of correctness of claims. Finally, we evaluated the performance of the proposed algorithms on real-world data sets collected from Twitter. The evaluation results demonstrated that the proposed algorithms can improve the estimation accuracy compared to the baselines. In addition, the PEM-MultiF algorithm has a higher estimation accuracy than the EM-MultiF algorithm in the absence of malicious sources. CEM-MultiF and EM-MultiF perform better than the existing truth-finding algorithms. What is more, CEM-MultiF has the higher recall but worse precision than PEM-MultiF.

Acknowledgments

Research reported in this article was sponsored in part by DARPA award W911NF-17-C-0099, DTRA award HDTRA1-18-1-0026, and the Army Research Laboratory under Cooperative Agreements W911NF-09-2-0053 and W911NF-17-2-0196. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies of the CCDC Army Research Laboratory or the US government. The US government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation hereon.

References

[1] Y. Li *et al.*, "On the discovery of evolving truth," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2015, pp. 675–684.

[2] S. Wang *et al.*, "Scalable social sensing of interdependent phenomena," in *Proc. 14th Int. Conf. Inf. Process. Sensor Netw.*, 2015, pp. 202–213.

[3] S. Yao *et al.*, "Recursive ground truth estimator for social data streams," in *Proc. 15th ACM/IEEE Int. Conf. Inf. Process. Sensor Netw.*, 2016, pp. 1–12.

[4] D. Wang *et al.*, "Using humans as sensors: An estimation-theoretic perspective," in *Proc. 13th Int. Symp. Inf. Process. Sensor Netw.*, 2014, pp. 35–46.

[5] X. Dong *et al.*, "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2014, pp. 601–610.

[6] M. Jiang *et al.*, "MetaPAD: Meta pattern discovery from massive text corpora," 2017, *arXiv: 1703.04213*.

[7] S. Liu, Z. Zheng, F. Wu, S. Tang, and G. Chen, "Context-aware data quality estimation in mobile crowdsensing," in *Proc. IEEE Conf. Comput. Commun.*, 2017, pp. 1–9.

[8] H. Zhang, Y. Li, F. Ma, J. Gao, and L. Su, "TextTruth: An unsupervised approach to discover trustworthy information from multi-sourced text data," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 2729–2737.

[9] X. Gong and N. Shroff, "Incentivizing truthful data quality for quality-aware mobile data crowdsourcing," in *Proc. 18th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2018, pp. 161–170.

[10] D. Zhou, J. He, K. S. Candan, and H. Davulcu, "MUVIR: Multi-view rare category detection," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 4098–4104.

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

[12] Y. Zhou and J. He, "Crowdsourcing via tensor augmentation and completion," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 2435–2441.

[13] Y. Zhou, A. R. Nelakurthi, and J. He, "Unlearn what you have learned: Adaptive crowd teaching with exponentially decayed memory learners," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 2817–2826.

[14] H. Shao *et al.*, "A constrained maximum likelihood estimator for unguided social sensing," in *Proc. IEEE Conf. Comput. Commun.*, 2018, pp. 2429–2437.

[15] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 675–684.

[16] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on Twitter," in *Proc. 7th Annu. Collaboration Electron. Messaging Anti-Abuse Spam Conf.*, 2010, vol. 6, Art. no. 12.

[17] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "TweetCred: Real-time credibility assessment of content on Twitter," in *Proc. Int. Conf. Soc. Informat.*, 2014, pp. 228–243.

[18] H. Shao *et al.*, "Unsupervised fact-finding with multi-modal data in social sensing," in *Proc. 22nd Int. Conf. Inf. Fusion*, 2019, pp. 1–8.

[19] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proc. Int. Conf. Web Search Data Mining*, 2008, pp. 183–194.

[20] A. Huang, "Similarity measures for text document clustering," in *Proc. 6th New Zealand Comput. Sci. Res. Student Conf.*, 2008, vol. 4, pp. 9–56.

[21] 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. Conf. Inf. Process. Sensor Netw.*, 2012, pp. 233–244.

[22] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 6, pp. 796–808, Jun. 2008.

[23] B. Zhao, B. I. Rubinstein, J. Gemmell, and J. Han, "A Bayesian approach to discovering truth from conflicting sources for data integration," *Proc. VLDB Endowment*, vol. 5, no. 6, pp. 550–561, 2012.

[24] D. Wang, T. Abdelzaher, L. Kaplan, R. Ganti, S. Hu, and H. Liu, "Exploitation of physical constraints for reliable social sensing," in *Proc. IEEE Real-Time Syst. Symp.*, 2013, pp. 212–223.

[25] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Integrating conflicting data: The role of source dependence," *Proc. VLDB Endowment*, vol. 2, no. 1, pp. 550–561, 2009.

[26] K. D. Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking, "Topical clustering of tweets," *Proc. ACM SIGIR: Workshop Soc. Web Search Mining*, 2011.

[27] H. Xiao, J. Gao, Z. Wang, S. Wang, L. Su, and H. Liu, "A truth discovery approach with theoretical guarantee," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1925–1934.

[28] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newslett.*, vol. 19, no. 1, pp. 22–36, 2017.

[29] L. Ma, W. P. Tay, and G. Xiao, "Iterative expectation maximization for reliable social sensing with information flows," *Inf. Sci.*, vol. 501, pp. 621–634, 2019.

[30] H. Shao *et al.*, "Optimizing source selection in social sensing in the presence of influence graphs," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst.*, 2017, pp. 1157–1167.

[31] H. Amintoosi, S. S. Kanhere, and M. Allahbakhsh, "Trust-based privacy-aware participant selection in social participatory sensing," *J. Inf. Secur. Appl.*, vol. 20, pp. 11–25, 2015.

[32] F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on sina weibo," in *Proc. ACM SIGKDD Workshop Mining Data Semantics*, 2012, Art. no. 13.

[33] J. Ma *et al.*, "Detecting rumors from microblogs with recurrent neural networks," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 3818–3824.

[34] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *Proc. IEEE 13th Int. Conf. Data Mining*, 2013, pp. 1103–1108.

[35] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: Characterizing and identifying fake images on Twitter during hurricane Sandy," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 729–736.

[36] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," 2019, *arXiv: 1902.06673*.

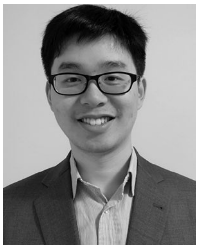
[37] D. Zhou, J. He, H. Yang, and W. Fan, "SPARC: Self-paced network representation for few-shot rare category characterization," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 2807–2816.

[38] H. Shao *et al.*, "paper2repo: GitHub repository recommendation for academic papers," in *Proc. Web Conf.*, 2020, pp. 629–639.

[39] A. Benamira, B. Devillers, E. Lesot, A. K. Ray, M. Saadi, and F. D. Malliaros, "Semi-supervised learning and graph neural networks for fake news detection," in *Proc. IEEE/ACM Int. Conf. Advances Soc. Netw. Anal. Mining*, 2019, pp. 568–569.

Huajie Shao received the BS degree from Jiangnan University, China, in 2011, and the MS degree from Zhejiang University, China, in 2014. He is currently working toward the PhD degree in computer science at the University of Illinois at Urbana Champagin (UIUC), Champaign, Illinois. His research interests mainly focus on data mining, deep learning-based recommender system, and social sensing. He has received ICCPS'17 Best Paper Award and FUSION'19 Best Student Paper Award.

Dachun Sun is currently working toward the senior undergraduate degree of computer science at the University of Illinois at Urbana Champagin (UIUC), Champaign, Illinois. He has a background of machine learning and software engineering. His research interest lies in computer vision and computational network analysis.



Shuochao Yao received the PhD degree from the Department of Computer Science, University of Illinois Urbana-Champaign, Champaign, Illinois. He is a postdoc researcher with the Department of Computer Science, University of Illinois Urbana-Champaign (UIUC), Champaign, Illinois. His research interests include deep learning on Internet of Things (IoT), cyber-physical systems, and crowd and social sensing.

Lu Su received the MS degree in statistics and PhD degree in computer science, both from the University of Illinois at Urbana-Champaign, Champaign, Illinois, in 2012 and 2013, respectively. He is an assistant professor with the Department of Computer Science and Engineering, SUNY Buffalo, Buffalo, New York. His research focuses on the general areas of mobile and crowd sensing systems, Internet of Things, and cyber-physical systems. He is the recipient of NSF CAREER Award, University at Buffalo Young Investigator Award.

Zhibo Wang received the BE degree in automation from Zhejiang University, China, in 2007, and the PhD degree in electrical engineering and computer science from the University of Tennessee, Knoxville, Tennessee, in 2014. He is currently an associate professor with the School of Cyber Science and Engineering, Wuhan University, China. His currently research interests include mobile crowdsensing systems, cyber-physical systems, recommender systems, and privacy protection.

Dongxin Liu received the BS and MS degrees in computer science from Shanghai Jiao Tong University, China, in 2014 and 2017, respectively. He is currently working toward the PhD degree with Computer Science Department, University of Illinois at Urbana-Champaign, Champaign, Illinois. His research interests encompass indoor white space exploration, social data analysis, and deep learning in IoT.

Shengzhong Liu received the BSc degree in computer science from Shanghai Jiao Tong University, China, in 2017. He is currently working toward the third-year PhD degree with the Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, Illinois. His research broadly lies in the intersection of social network analysis, social sensing, and machine learning.

Lance Kaplan (Fellow, IEEE) received the BS degree with distinction from Duke University, Durham, North Carolina, in 1989 and the MS and PhD degrees from the University of Southern California, Los Angeles, California, in 1991 and 1994, respectively, all in electrical engineering. From 1996–2004, he was a member of the faculty with the Department of Engineering and a senior investigator with the Center of Theoretical Studies of Physical Systems (CTSPS) at Clark Atlanta University (CAU), Atlanta, Georgia. Currently, he is a researcher with the Context Aware Processing branch of the U.S Army Research Laboratory (ARL). He serves on the Board of Governors for the IEEE Aerospace and Electronic Systems (AES) Society (2008-2013, 2018-Present) and as VP of Conferences for the International Society of Information Fusion (ISIF) (2014-Present). Previously, he served as an editor-In-chief for the *IEEE Transactions on Aerospace and Electronic Systems* (2012–2017) and on the board of directors of ISIF (2012–2014). He is a three time recipient of the Clark Atlanta University Electrical Engineering Instructional Excellence Award from 1999–2001. He is a fellow of ARL. His current research interests include information/data fusion, reasoning under uncertainty, network science, resource management and signal, and image processing.

Tarek Abdelzaher (Senior Member, IEEE) is currently a professor and willett faculty scholar with the Department of Computer Science, the University of Illinois at Urbana Champaign, Champaign, Illinois. He has authored/coauthored more than 300 refereed publications in real-time computing, distributed systems, sensor networks, and control. He served as an editor-in-chief of the *Journal of Real-Time Systems*, and as associate editor of multiple journals including the *IEEE Transactions on Mobile Computing*, *IEEE Transactions on Parallel and Distributed Systems*, *ACM Transaction on Sensor Networks*, *ACM Transactions on Internet of Things*,*ACM Transactions on Internet Technology*, and *Ad Hoc Networks Journal*, among others. His research interests lie broadly in understanding and influencing performance and temporal properties of networked embedded, social and software systems in the face of increasing complexity, distribution, and degree of interaction with an external physical environment. He is a fellow of ACM.