Iterative expectation maximization for reliable social sensing with information flows

Lijia Ma*, Wee Peng Tay, and Gaoxi Xiao

School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore

Abstract

Social sensing relies on a large number of observations reported by different, possibly unreliable, agents to determine if an event has occurred or not. In this paper, we consider the truth discovery problem in social sensing, in which an agent may receive another agent's observation (known as an information flow), and may change its observation to match the observation it receives. If an agent's observation is influenced by another agent, we say that the former is a dependent agent. We propose an Iterative Expectation Maximization algorithm for Truth Discovery (IEMTD) in social sensing with dependent agents. Compared with other popular truth discovery approaches, which assume either the agents' observations are independent, or their dependency is known a priori, IEMTD allows to infer each agent's reliability, the observations' dependency and the events' truth jointly. Simulation results on synthetic data and three real world data sets demonstrate that in almost all our experiments, IEMTD achieves a higher truth discovery accuracy than the existing algorithms when dependencies exist between agents' observations.

Keywords: Truth discovery, social sensing, reliability, information flow, expectation maximization.

1. Introduction

With the rapid development of communications, it is becoming increasingly more convenient for agents to report observations about the state of events in the physical world. For instance, drivers may report traffic congestion information to traffic control centers using mobile phones, and city dwellers may report sources of noise pollution, littering and suspicious activities to their local authorities via online media platforms like Facebook and Twitter. Social sensing has become a popular sensing paradigm. In social sensing, observations from agents are collected and aggregated to perform inference [1, 11, 19, 22, 29]. This paradigm has seen many interesting applications, including the truth discovery, crowdsourcing, copy detection, semantic-aware recommendation, event detection, traffic congestion detection, rumor identification, disaster estimation and image classification [2, 5, 6, 8, 12, 21, 28].

In the truth discovery problem, we aim to find the true state of events from the collected observations [14, 32]. In this paper, we consider the case where the state of an event is either true or false, which may correspond to whether the event has happened or not, respectively. Moreover, agents' observations are assumed to be binary. Examples of such events include whether a traffic accident has occurred along a particular road, and whether a shooting incident has happened in a particular area [22, 29]. Inferring the truth of an event is challenging as agents' observations conflict with each other generally, and they are easily influenced by other agents with whom they are connected via a social network [7, 10, 26, 31]. For instance, a post on Facebook regarding an event may be strongly influenced by another post, although it may appear to a reader to be an independent post.

Several truth discovery methods have been proposed, most of which focus on estimating the reliability of agents and assuming that agents' observations are independent. Here, the reliability of an agent denotes the probability that the agent reports a correct observation [29]. Classical methods include the majority voting (MV), optimization [14, 16] and maximum likelihood estimation (MLE) methods [29, 34]. The MV methods assume that agents have the same reliability, and infer the truth of events as that returned by the majority of agents [33]. The optimization methods [14, 16] infer the truth of events by minimizing the weighted deviation between the agents' observations and the inferred state of events, in which the weight represents the reliability of agents. The MLE methods [29, 34] use a vector of confusion parameters to represent the reliability of an agent, and infer the truth of events and the reliability of agents using an expectation maximization (EM) algorithm. These methods were generalized and extended by introducing various fact confidences, agent credibility, agent sociability, agent mobility, event correlation and agents' response times [8, 9, 13, 15, 18, 20, 21, 23, 27, 30, 32]. The above-mentioned methods assume that all agents' observations are independent. In some social sensing applications, agents can be influenced by the observations of other agents. Examples include retweets and repostings in Twitter, in which agents may repost other agents' messages as their own observations [25]. We say that there is an information flow

*Corresponding author
Email address: omegamalj@gmail.com (Lijia Ma)

from one agent to another if the former shares its observation or other information to the latter. As a result, the latter agent may be influenced to change its observation to match that of the former agent. In this case, we say that the latter agent is dependent on the former agent. To effectively infer the truth of events in such applications, the dependency between agents' observations needs to be considered.

Existing approaches typically assume that the dependency between agents' observations is known a priori. The information copy approach is widely used to model this dependency [4, 7, 26]: if two agents make many common observations, their observations are probably dependent on each other. MalVoteCount [4], Apollo-social [26] and TA-EM [7] adopt the information copy approach to describe the dependency between agents' observations. Moreover, MalVoteCount [4] adopts a Bayesian analysis to determine the dependency degree between agents' observations and the truth of events while Apollo-social [26] and TA-EM [7] use MLE based methods to estimate the reliability of agents and the truth of events. In some applications, the information copy approach cannot correctly model the dependency between agents' observations as any two agents with high reliability may share a large number of identical observations. Examples include asking movie critics in different areas to rate movies, or asking drivers in the same area who do not know each other to report the traffic congestion of roads. Although the observations of agents are independent of each other, they may give the same observations.

In some literature, the dependency between agents' observations is modeled based on their social connections, e.g., friendships, online communications, or physical connections. The dependency based expectation maximization (DEM) technique [31] uses information flows to represent the social correlations, which stem from communications, e.g., drivers can propagate the traffic congestion information of roads to other nearby drivers using wireless communications, and rumors can propagate information on WeChat, Facebook and Twitter platforms. To infer the dependency between agents' observations, DEM assumes that an agent's observation is dependent on those other agents from whom there are information flows and they share the same observation.

In this work, we study the truth discovery problem in social sensing, in which there exist information flows among agents, the observations of agents maybe dependent and their dependency is unknown a priori. We propose an iterative EM truth discovery (IEMTD) algorithm that, similar to DEM, uses the information flows to represent the social correlations between agents. Different from DEM, however, IEMTD infers the unknown dependency between agents' observations based on a maximum likelihood estimation approach. Our main contributions are as follows:

1. We develop a truth discovery model that takes into account i) the reliability of agents; ii) the unknown dependency between agents' observations; and iii) the interplay between agents' reliability and dependency.

2. We propose the IEMTD algorithm to infer i) the truth of events; ii) the reliability of agents; and iii) the unknown dependency between agents' observations.

3. We verify the performance of the proposed algorithm through extensive experiments on both synthetic and three real world data sets. The experiments suggest that when the agents' observations are dependent, IEMTD outperforms the current state-of-the-art truth discovery approaches, which either assume that the agents'observations are independent or have a specific known dependency structure.

The rest of the paper is organized as follows. In Section 2, we provide our truth discovery system model and problem definition. In Section 3, we introduce the details of our algorithm. Simulations and experiments on real data are presented in Section 4. Finally, we conclude and briefly discuss future work in Section 5.

Notations: We use bold faced symbols to represent random variables. The calligraphic letters are used to denote sets. Let A be a matrix, and A_{ij} be the (i,j)th entry of A. We use $\cdot$ in place of i or j if the index set spans all the row or column indices of A, respectively. To avoid cluttered notations, we use $p(S)$ to denote the probability mass function of the discrete random variable $\boldsymbol{S}$ evaluated at $\boldsymbol{S} = S$. The same convention applies for conditional probabilities, e.g., $p(S \mid \mathrm{A})$. The operator $\mathbb{E}$ denotes mathematical expectation. The notation $X \sim \mathrm{Bern}\,(p)$ means that the binary random variable X follows a Bernoulli distribution with probability p. We use the operators $|S|$ and $\|S\|$ to denote the number of elements and the sum of the absolute value of elements in S, respectively.

2. System model and problem formulation

In this section, the truth discovery system model is first introduced, and then the problem formulation is given.

2.1. System model

We consider a social sensing system model with m events $\mathcal{E} = \{1, 2, \ldots, m\}$, and q agents $\mathcal{J} = \{1, 2, \ldots, q\}$ (shown in 1). Each event $i \in \mathcal{E}$ has an associated true state $\boldsymbol{S}_i \in \{0,\ 1\}$ denoting that the event i has happened ($\boldsymbol{S}_i = 1$) or not

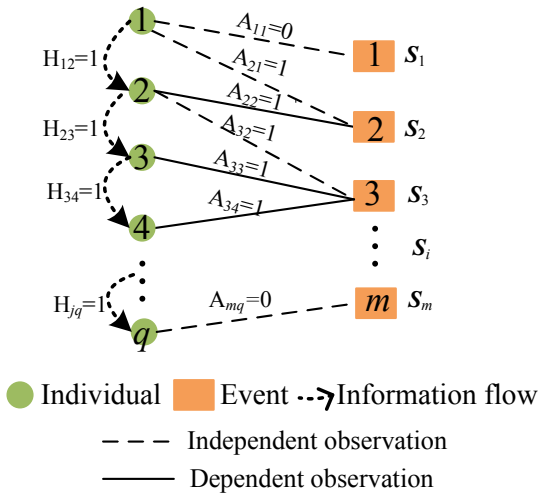


Figure 1: Our truth discovery system model.

($\boldsymbol{S}_i = 0$), and it can be observed by multiple agents. Each agent reports a claim about each event it has observed to a central decision maker. For each event i and agent j, we let

$$\mathrm{A}_{ij} = \begin{cases} -\infty & \text{if event } i \text{ is not observed by agent } j, \\ 1 & \text{if event } i \text{ is observed by agent } j \text{ who claims that the event has happened,} \\ 0 & \text{if event } i \text{ is observed by agent } j \text{ who claims that the event has not happened.} \end{cases} \tag{1}$$

Let $\mathrm{A} = [\mathrm{A}_{ij}]_{i\in\mathcal{E}, j\in\mathcal{J}}$. In this paper, we use the terms "claim" and "observation", "individual" and "agent" interchangeably. The main features of our system model are as follows. See Fig. 1 for an illustration.

1) For each event $i \in \mathcal{E}$, we let its true state $\boldsymbol{S}_i \sim \mathrm{Bern}\,(p_s)$, where p_s is the probability that the event has occurred. We assume that events are independent of each other. Let $\boldsymbol{S} = \{\boldsymbol{S}_i : i \in \mathcal{E}\}$.

2) An agent may propagate its claim to other agents. If there is an information flow from agent k to agent j, we say that agent k is an ancestor of agent j, and let $\mathrm{H}_{kj} = 1$. If there is no information flow, $\mathrm{H}_{kj} = 0$. Let $\mathrm{H} = [\mathrm{H}_{kj}]_{k,j\in\mathcal{J}}$.

3) Once an agent receives information from its ancestors, its claims may or may not be influenced by the received information. Here, we consider that each agent j may be influenced by his ancestors who give the same observations as his observations to the same events. For each agent j and its observed event i, we let $\mathcal{K}_{ij} = \{k \in \mathcal{J} : \mathrm{H}_{kj} = 1, \mathrm{A}_{ik} = \mathrm{A}_{ij}\}$ be the set of individual j' ancestors who give the same observations as j to event i. Let $\mathcal{K} = \{\mathcal{K}_{ij}, i \in \mathcal{E}, j \in \mathcal{J}\}$. We use $\mathbf{D}_{ij} = 1$ to denote the case where agent j's claim about event i has been influenced by its ancestors, and $\mathbf{D}_{ij} = 0$ to represent otherwise. For each $a \in \{0,\ 1\}$, we model

$$\mathbf{D}_{ij} = \begin{cases} 0 & \text{if } |\mathcal{K}_{ij}| = 0, \\ \boldsymbol{C}_j^a & \text{if } |\mathcal{K}_{ij}| \neq 0 \text{ and } \boldsymbol{S}_i = a, \end{cases} \tag{2}$$

where $\boldsymbol{C}_j^a \sim \mathrm{Bern}\,(p_d^a)$ is a Bernoulli random variable, which denotes the truth dependency for an individual to give observations about an event i with $\boldsymbol{S}_i = a$. The probability $1-p_d^a$ represents the chance that an agent is familiar with an event i with $\boldsymbol{S}_i = a$, e.g., it may have previously observed the same event or have prior knowledge about the event. An example is the case where traffic jams on a particular segment of a road may happen very frequently, therefore an agent j who has experienced this event multiple times previously will not be easily influenced by the opinions of its ancestors. We call $p_d = \{p_d^a : a \in \{0,\ 1\}\}$ the probability for an individual to give dependent observations to an event that has (not) happened with $a = 1$ ($a = 0$). The p_d values are unknown a priori and are estimated by our IEMTD method.

Note that in our model, for all $i \in \mathcal{E}$ such that $|\mathcal{K}_{ij}| \neq 0$, their corresponding $\mathbf{D}_{ij}$ are coupled through $\boldsymbol{C}_j^a$, $a \in \{0,\ 1\}$. This models the scenario where if an agent j is influenced in one event, then it is also likely to be influenced in all other events with the same state. We assume that $\{\boldsymbol{C}_j^a : j \in \mathcal{J}, a = 0, 1\}$ are independent.

Let $\mathbf{D} = [\mathbf{D}_{ij}]_{i\in\mathcal{E}, j\in\mathcal{J}}$. Note that our dependency matrix $\mathbf{D}$ is different from the dependency definition in [31], which defines $\mathbf{D}_{ij}$ to be whether the observation A_{ij} by agent j is the same as its ancestors' observation, and assumes that this dependency is known a priori. In this paper, $\mathbf{D}$ is a random matrix to be inferred.

4) For each agent j, and $a, b \in \{0,\ 1\}$, let $R_j^{ab} = p(\mathrm{A}_{ij} = a \mid \boldsymbol{S}_i = a, \mathbf{D}_{ij} = b)$ be the probability that agent j reports a correct observation to an event i conditioned on $\boldsymbol{S}_i = a$ and $\mathbf{D}_{ij} = b$. We call $R_j = \{R_j^{ab} : a, b \in \{0,\ 1\}\}$ the reliability of agent j. The R_j values are unknown a priori and are estimated by our IEMTD method. Our reliability R_j formulation can be considered as a generalized version of that in [31], with $p_d^1 = p_d^0 = 1$ and $R_j^{10} = a_j, R_j^{11} = f_j, R_j^{00} = 1 - b_j, R_j^{01} = 1 - g_j$, where we refer the reader to [31] for the definitions of the parameters a_j, b_j, f_j, g_j.

5) Let $\mathcal{J}^0 = \{j \in \mathcal{J} : \mathrm{H}_{kj} = 0, \forall k \in \mathcal{J}\}$ be the set of agents who have not received information from other agents and let $\mathcal{J}^1$ denote the set of agents with ancestors. If event i is observed by agent $j \in \mathcal{J}^0$, we have

$$p(\mathrm{A}_{ij} \mid S_i; \mathrm{H}_{\cdot j}, R_j) = \begin{cases} (R_j^{10})^{\mathrm{A}_{ij}} \cdot (1 - R_j^{10})^{(1-\mathrm{A}_{ij})} & \text{if } S_i = 1, \\ (R_j^{00})^{(1-\mathrm{A}_{ij})} \cdot (1 - R_j^{00})^{\mathrm{A}_{ij}} & \text{if } S_i = 0. \end{cases}$$

For an event i observed by an agent $j \in \mathcal{J}^1$, it can be shown that

$$p(\mathrm{A}_{ij} \mid \mathcal{K}_{ij}, S_i, \mathrm{D}_{ij}; \mathrm{H}_{\cdot j}, R_j) = \begin{cases} (R_j^{10})^{\mathrm{A}_{ij}} \cdot (1 - R_j^{10})^{(1-\mathrm{A}_{ij})} & \text{if } |\mathcal{K}_{ij}| = 0,\ S_i = 1 \text{ and } \mathrm{D}_{ij} = 0, \\ (R_j^{00})^{(1-\mathrm{A}_{ij})} \cdot (1 - R_j^{00})^{\mathrm{A}_{ij}} & \text{if } |\mathcal{K}_{ij}| = 0,\ S_i = 0 \text{ and } \mathrm{D}_{ij} = 0, \\ (R_j^{10})^{\mathrm{A}_{ij}} \cdot (1 - R_j^{10})^{(1-\mathrm{A}_{ij})} & \text{if } |\mathcal{K}_{ij}| \neq 0,\ S_i = 1 \text{ and } \mathrm{D}_{ij} = 0, \\ (R_j^{11})^{\mathrm{A}_{ij}} \cdot (1 - R_j^{11})^{(1-\mathrm{A}_{ij})} & \text{if } |\mathcal{K}_{ij}| \neq 0,\ S_i = 1 \text{ and } \mathrm{D}_{ij} = 1, \\ (R_j^{01})^{(1-\mathrm{A}_{ij})} \cdot (1 - R_j^{01})^{\mathrm{A}_{ij}} & \text{if } |\mathcal{K}_{ij}| \neq 0,\ S_i = 0 \text{ and } \mathrm{D}_{ij} = 1, \\ (R_j^{00})^{(1-\mathrm{A}_{ij})} \cdot (1 - R_j^{00})^{\mathrm{A}_{ij}} & \text{if } |\mathcal{K}_{ij}| \neq 0,\ S_i = 0 \text{ and } \mathrm{D}_{ij} = 0. \end{cases}$$

We assume that for each $i \in \mathcal{E}$, conditioned on $\boldsymbol{S}_i$ and $\{\mathbf{D}_{ij}, \mathcal{K}_{ij} : j \in \mathcal{J}\}$, $\{\mathrm{A}_{ij} : j \in \mathcal{J}\}$ are independent.

The commonly used symbols in our system model are summarized in Table 1 for easy reference.

Table 1: Commonly used symbols.

Symbol	Definition
$\boldsymbol{S}_i$	The true state of event i.
A_{ij}	Claim of agent j about the state of event i; see Eq. (1).
H_{kj}	Indicator of whether there is information flow from agent k to agent j.
$\mathbf{D}_{ij}$	Dependency indicator of the claim of agent j about event i; see Eq. (2).
R_j	Reliability parameters $R_j = \{R_j^{ab} : a, b \in \{0,\ 1\}\}$ of agent j, with $R_j^{ab} = p(\mathrm{A}_{ij} = a \mid \boldsymbol{S}_i = a, \mathbf{D}_{ij} = b)$.
$\mathcal{K}_{ij}$	The set of individual j' ancestors who give the same observations as j to event i
m, q	Number of events and agents, respectively.
p_s	The probability that an event has happened.
p_d	The probability $p_d = \{p_d^a : a \in \{0,\ 1\}\}$ that an agent gives dependent observations for an event that has (not) happened with $a = 1$ ($a = 0$).
C_j	The familiarity $C_j = \{C_j^a : a \in \{0,\ 1\}\}$ of agent j to an event that has (not) happened with $a = 1$ ($a = 0$).
$\mathcal{J}^1(\mathcal{J}^0)$	The set of agents who have (not) received information from other agents.

2.2. Problem formulation

Given a truth discovery system with m events, q agents and information flows H, the objective of our truth discovery problem is to find the optimal estimators of unknown parameters $\Theta = \{R, p_s, p_d\}$, where $R = \{R_j : j \in \mathcal{J}\}$, and infer the states $\boldsymbol{S} = \{\boldsymbol{S}_i : i \in \mathcal{E}\}$ and $\mathbf{D}$, so as to maximize the marginal likelihood or probability of observations A. The marginal likelihood of A is given by

$$\begin{aligned} p(\mathrm{A}; \mathrm{H}, \Theta) &= \sum_{S, \mathrm{D}} p(\mathrm{A}, S, \mathrm{D}; \mathrm{H}, \Theta) \\ &= \sum_{S, \mathrm{D}} p(S, \mathrm{D}; \mathrm{H}) \cdot p(\mathrm{A} \mid S, \mathrm{D}; \mathrm{H}, \Theta) \\ &= \sum_{S, \mathrm{D}} p(S) \cdot p(\mathrm{D} \mid S; \mathrm{H}) \cdot p(\mathrm{A} \mid S, \mathrm{D}; \mathrm{H}, \Theta), \end{aligned} \tag{3}$$

where

$$p(S) = \prod_{i=1}^{m} p(S_i), \tag{4}$$

$$p(\mathrm{D} \mid S; \mathrm{H}) = \prod_{j \in \mathcal{J}^1} p(\mathrm{D}_{\cdot j} \mid S; \mathrm{H}_{\cdot j}) = \prod_{a \in \{1,\ 0\}} \prod_{j \in \mathcal{J}^1 : |\mathcal{K}_{ij}| \neq 0, S_i = a, \exists i \in \mathcal{E}} p(C_j^a), \tag{5}$$

$$p(\mathrm{A} \mid S, \mathrm{D}; \mathrm{H}, \Theta) = \prod_{i=1}^{m} \Big(\prod_{j \in \mathcal{J}^0} p(\mathrm{A}_{ij} \mid S_i; \mathrm{H}_{\cdot j}, R_j) \cdot \prod_{j \in \mathcal{J}^1} p(\mathrm{A}_{ij} \mid \mathcal{K}_{ij}, S_i, \mathrm{D}_{ij}; \mathrm{H}_{\cdot j}, R_j) \Big). \tag{6}$$

We seek to solve the following optimization problems:

$$\begin{aligned} \Theta^* &\leftarrow \arg\max_{\Theta} p(\mathrm{A}; \mathrm{H}, \Theta), \\ (S^*, \mathrm{D}^*) &\leftarrow \arg\max_{(S, \mathrm{D})} p(S, \mathrm{D} \mid \mathrm{A}; \mathrm{H}, \Theta^*). \end{aligned} \tag{7}$$

In the next section, we present an iterative optimization algorithm for solving the problem in Eq. (7).

3. Our solution

Solving the optimization problem in Eq. (7) is computationally hard as it involves the combinatorial optimization of both Θ, S and D. A single EM algorithm [3, 24] treating both D and S as latent variables at the same time needs to estimate the joint distributions of D and S. Such an approach's computational complexity is exponentially increasing with the number of active ancestors for each agent, the number of events for each agent and the number of agents for each event. Therefore, in this section, we present an iterative EM algorithm (which we call IEMTD) to estimate Θ by alternatively treating D or S as latent variables.

Algorithm 1 Framework of our IEMTD

Input: A and t_{max}.
Output: D^*, Θ^* and S^*.
$t \leftarrow 0$.
$S^{(0)}, \mathrm{D}^{(0)} \leftarrow$ initialize S and D based on majority voting, respectively.
$\Theta^{(0)} \leftarrow$ compute the parameters in Θ by Eq. (8) based on the initialized $S^{(0)}$ and $\mathrm{D}^{(0)}$.
while $||\mathrm{D}^{(t+1)} - \mathrm{D}^{(t)}|| + ||S^{(t+1)} - S^{(t)}|| + ||\Theta^{(t+1)} - \Theta^{(t)}|| \geq 0.01$ and $t < t_{max}$ **do**
    $(S^{(t+1)}, \Theta^{(t+1)}) \leftarrow$ **EMS**$(\mathrm{A}, \mathrm{D}^{(t)}, \Theta^{(t)})$.
    $(\mathrm{D}^{(t+1)}, \Theta^{(t+1)}) \leftarrow$ **EMD**$(\mathrm{A}, S^{(t+1)}, \Theta^{(t+1)})$.
    $t \leftarrow t + 1$.
end while
$\mathrm{D}^* \leftarrow \mathrm{D}^{(t)}$, $\Theta^* \leftarrow \Theta^{(t)}$ and $S^* \leftarrow S^{(t)}$.

The IEMTD algorithm alternates between the following two optimization steps: i) optimization of Θ together with the most likely S and a fixed D, and ii) optimization of Θ together with the most likely D and a fixed S. At each step, IEMTD uses an EM algorithm to find an optimal estimator of Θ together with the most likely value of S or D. IEMTD alternates between the two steps until it satisfies a convergence criterion: either the values of S and D remain constant, and the difference in norm of the estimated Θ is smaller than a given threshold value, or the number of iterations between subsequent iterations t reaches a maximum number $t_{\max}$. The main framework of IEMTD is given in **Algorithm 1**. In **Algorithm 1**, the functions **EMS** and **EMD** are the corresponding EM algorithms in the two optimization steps, which we discuss in detail in Section 3.2 and Section 3.3, respectively.

EM guarantees that the marginal likelihood of the observed data is non-decreasing over iterations [3, 24]. However, it cannot guarantee convergence to a maximum likelihood estimator. In other words, for some practical applications with multimodal distributions, the EM algorithm may fall into the local optimum. Our experiment results in Section 4 suggest that the IEMTD algorithm, which iteratively executes two EM algorithms (i.e., EMS and EMD) that optimize over S and D, respectively, is less likely to be trapped in a local optimum, when compared to the DEM algorithm [31].

3.1. Initialization

As the MV method has a low computational complexity and its solution is consistent with the majority of agents' observations, we use it to initialize S and D at the iteration $t = 0$. More specifically, for each event i, its initial $S_i^{(0)}$ value is the observation value returned by the majority of agents who observe event i. For each observation A_{ij}, the initial $\mathrm{D}_{ij}^{(0)}$ value is 1 if agent j gives the same observation as the majority of his ancestors, and 0 otherwise. Note that, there exist

Algorithm 2 EMS

Input: A, $\mathrm{D}^{(t)}$, $\Theta^{(t)}$ and t_{max}.
Output: $S^{(t+1)}$ and $\Theta^{(t+1)}$.
$n \leftarrow 0$, $\Theta^{(n)} \leftarrow \Theta^{(t)}$, $\mathrm{D} \leftarrow \mathrm{D}^{(t)}$.
while $||\Theta^{(n+1)} - \Theta^{(n)}|| \geq 0.01$ and $n < t_{max}$ **do**
    E-step: Compute the expectation of the log likelihood function $\mathbf{Q}(\Theta \mid \Theta^{(n)})$ of $p(\mathrm{A}, S \mid \mathrm{D}; \mathrm{H}, \Theta)$ based on Eq. (11).
    M-step: Maximize $\mathbf{Q}(\Theta \mid \Theta^{(n)})$, and then we get $\Theta^{(n+1)}$ based on Eqs. (12) and (13).
    $n \leftarrow n + 1$.
end while
$\Theta^{(t+1)} \leftarrow \Theta^{(n)}$.
Compute $S^{(t+1)}$ based on Eq. (14).

tie in which it is the same for the number of observations with '0' and '1' to the event i (the agent j is consistent with exactly half of his ancestors). In this case, $S_i^{(0)}$ ($\mathrm{D}_{ij}^{(0)}$) is randomly set as 0 or 1.

Using the initialized $S^{(0)}$ and $\mathrm{D}^{(0)}$, for each $j \in \mathcal{J}$ and $a, b \in \{0,\ 1\}$, we can initialize the parameters in Θ as follows:

$$\begin{aligned}
(R_j^{ab})^{(0)} &= \frac{|\{i \in \mathcal{E} : \mathrm{A}_{ij} = S_i^{(0)}, S_i^{(0)} = a, \mathrm{D}_{ij}^{(0)} = b\}|}{|\{i \in \mathcal{E} : S_i^{(0)} = a, \mathrm{D}_{ij}^{(0)} = b\}|}, \\
p_s^{(0)} &= \frac{|\{i \in \mathcal{E} : S_i^{(0)} = 1\}|}{m}, \\
(p_d^a)^{(0)} &= \frac{|\{j \in J^1 : |\mathcal{K}_{ij}| \neq 0, S_i^{(0)} = a, D_{ij}^{(0)} = 1, \exists i \in \mathcal{E}\}|}{|\{j \in J^1 : |\mathcal{K}_{ij}| \neq 0, S_i^{(0)} = a, \exists i \in \mathcal{E}\}|}.
\end{aligned} \tag{8}$$

3.2. Finding $\Theta^{(t+1)}$ and $S^{(t+1)}$ with a fixed $\mathrm{D} = \mathrm{D}^{(t)}$

The problem in Eq. (7) at iteration t with a given $\mathrm{D} = \mathrm{D}^{(t)}$ is to find the optimal estimators of unknown parameters Θ together with the most likely truth S for events at iteration $t+1$, so as to maximize the probability of observations A, and it can be formulated as follows:

$$\begin{aligned}
\Theta^{(t+1)} &\leftarrow \arg\max_{\Theta} p(\mathrm{A} \mid \mathrm{D}; \mathrm{H}, \Theta), \\
S^{(t+1)} &\leftarrow \arg\max_{S} p(S \mid \mathrm{A}, \mathrm{D}; \mathrm{H}, \Theta^{(t+1)}),
\end{aligned} \tag{9}$$

where $p(\mathrm{A} \mid \mathrm{D}; \mathrm{H}, \Theta) = \sum_S p(\mathrm{A}, S \mid \mathrm{D}; \mathrm{H}, \Theta)$. Here, $\Theta = \{R, p_s, p_d\}$, and the parameters R and p_s need to be updated when D is given.

In Eq. (9), the optimal estimate of Θ is related to the marginal likelihood $p(\mathrm{A} \mid \mathrm{D}; \mathrm{H}, \Theta)$ of A. However, computing $p(\mathrm{A} \mid \mathrm{D}; \mathrm{H}, \Theta)$ is often intractable, which considers all possible values of $S = \{S_1, S_2, \ldots, S_q\}$. Here, an EM algorithm (called as EMS), which is to find the MLE of the unknown parameters in Θ for the statistic model in Eq. (9) with the observed data A and latent variables S, is presented to solve this problem. The corresponding likelihood function $p(\mathrm{A}, S \mid \mathrm{D}; \mathrm{H}, \Theta)$ is given as follows:

$$\begin{aligned}
p(\mathrm{A}, S \mid \mathrm{D}; \mathrm{H}, \Theta) &= p(\mathrm{A} \mid S, \mathrm{D}; \mathrm{H}, \Theta) \cdot p(S) \\
&= \prod_{i=1}^{m} p(\mathrm{A}_{i\cdot} \mid S_i, \mathrm{D}_{i\cdot}; \mathrm{H}, \Theta) \cdot p(S_i),
\end{aligned} \tag{10}$$

where $\mathrm{A}_{i\cdot}$ denotes i-th row of elements of A,

$$p(\mathrm{A}_{i\cdot} \mid S_i, \mathrm{D}_{i\cdot}; \mathrm{H}, \Theta) = \prod_{j \in \mathcal{J}^0} p(\mathrm{A}_{ij} \mid S_i; \mathrm{H}_{\cdot j}, R_j) \cdot \prod_{j \in \mathcal{J}^1} p(\mathrm{A}_{ij} \mid \mathcal{K}_{ij}, S_i, \mathrm{D}_{ij}; \mathrm{H}_{\cdot j}, R_j).$$

Next, the EMS algorithm maximizes this likelihood to estimate the parameters in Θ by iterating the following two steps: **E-step** and **M-step**.

E-step: Compute the expected value of the log likelihood $\mathbf{Q}(\Theta \mid \Theta^{(n)})$ of $p(\mathrm{A}, S \mid \mathrm{D}; \mathrm{H}, \Theta)$.

$$\begin{aligned}\mathbf{Q}(\Theta \mid \Theta^{(n)}) &= \mathbb{E}_{\boldsymbol{S}|\Theta^{(n)}}\Big[\ln p(\mathrm{A}, \boldsymbol{S} \mid \mathrm{D}; \mathrm{H}, \Theta)\Big] \\ &= \sum_{i=1}^{m} \sum_{a \in \{1,\ 0\}} p(S_i = a \mid \mathrm{A}_{i.}, \mathrm{D}_{i.}; \mathrm{H}, \Theta^{(n)}) \cdot \ln\Big(p(\mathrm{A}_{i.} \mid S_i = a, \mathrm{D}_{i.}; \mathrm{H}, \Theta) \cdot (p_s)^a \cdot (1-p_s)^{(1-a)}\Big).\end{aligned} \tag{11}$$

Here, $p(S_i = 1 \mid \mathrm{A}_{i.}, \mathrm{D}_{i.}; \mathrm{H}, \Theta^{(n)})$ $\Big(p(S_i = 0 \mid \mathrm{A}_{i.}, \mathrm{D}_{i.}; \mathrm{H}, \Theta^{(n)})\Big)$ is the conditional probability that the event i has (not) occurred given $\mathrm{A}_{i.}$ and $\mathrm{D}_{i.}$ under the current estimation of $\Theta^{(n)}$, and it can be computed as follows:

$$\begin{aligned}p(S_i = 1 \mid \mathrm{A}_{i.}, \mathrm{D}_{i.}; \mathrm{H}, \Theta^{(n)}) &= \frac{p(S_i = 1 \mid \mathrm{A}_{i.}, \mathrm{D}_{i.}; \mathrm{H}, \Theta^{(n)})}{\sum_{S_i \in \{0,\ 1\}} p(S_i \mid \mathrm{A}_{i.}, \mathrm{D}_{i.}; \mathrm{H}, \Theta^{(n)})} \\ &= \frac{p(S_i = 1) \cdot p(\mathrm{A}_{i.} \mid S_i = 1, \mathrm{D}_{i.}; \mathrm{H}, \Theta^{(n)})}{\sum_{S_i \in \{0,\ 1\}} p(S_i) \cdot p(\mathrm{A}_{i.} \mid S_i, \mathrm{D}_{i.}; \mathrm{H}, \Theta^{(n)})} \\ &= \frac{p_s \cdot p(\mathrm{A}_{i.} \mid S_i = 1, \mathrm{D}_{i.}; \mathrm{H}, \Theta^{(n)})}{p_s \cdot p(\mathrm{A}_{i.} \mid S_i = 1, \mathrm{D}_{i.}; \mathrm{H}, \Theta^{(n)}) + (1-p_s) \cdot p(\mathrm{A}_{i.} \mid S_i = 0, \mathrm{D}_{i.}; \mathrm{H}, \Theta^{(n)})},\end{aligned} \tag{12}$$

and $p(S_i = 0 \mid \mathrm{A}_{i.}, \mathrm{D}_{i.}; \mathrm{H}, \Theta^{(n)}) = 1 - p(S_i = 1 \mid \mathrm{A}_{i.}, \mathrm{D}_{i.}; \mathrm{H}, \Theta^{(n)})$.

M-step: Maximize the expected log function $\mathbf{Q}(\Theta \mid \Theta^{(n)})$, and then get the parameters in Θ of the next iteration. We take the derivatives of $\mathbf{Q}(\Theta \mid \Theta^{(n)})$ to each parameter in Θ, and find the optimal solutions when these derivatives are equal to 0. And thereafter, for each agent j, and $a, b \in \{0,\ 1\}$, we can get

$$\begin{aligned}(R_j^{ab})^{(n+1)} &= \frac{\sum_{i \in \mathcal{E}: \mathrm{A}_{ij} = a, \mathrm{D}_{ij} = b} p(S_i = a \mid \mathrm{A}_{i.}, \mathrm{D}_{i.}; \mathrm{H}, \Theta^{(n)})}{\sum_{i \in \mathcal{E}: \mathrm{D}_{ij} = b} p(S_i = a \mid \mathrm{A}_{i.}, \mathrm{D}_{i.}; \mathrm{H}, \Theta^{(n)})}, \\ p_s^{(n+1)} &= \frac{\sum_{i=1}^{m} p(S_i = 1 \mid \mathrm{A}_{i.}, \mathrm{D}_{i.}; \mathrm{H}, \Theta^{(n)})}{m}.\end{aligned} \tag{13}$$

The above E-step and M-step are iteratively executed until the estimated Θ values converge. Finally, the EMS algorithm evaluates the truth $S_i^{(t+1)}$ of each event $i \in \mathcal{E}$ as follows:

$$S_i^{(t+1)} \leftarrow \arg\max_{S_i \in \{0,\ 1\}} p(S_i \mid \mathrm{A}_{i.}, \mathrm{D}_{i.}; \mathrm{H}, \Theta^{(n+1)}). \tag{14}$$

The framework of the EMS algorithm is shown in **Algorithm 2**. More specifically, the inputs include the observations A, $\mathrm{D}^{(t)}$, $\Theta^{(t)}$ and maximum number of iterations t_{max}. Here, the values of $\mathrm{D}^{(t)}$ and $\Theta^{(t)}$ are computed by Eq. (8) when $t = 0$ and returned by the function EMD when $t > 0$. The EMS algorithm alternates between the E-step and M-step until it satisfies a convergence criterion: either the values of S remain constant, and the difference in norm of the estimated Θ between subsequent iterations is smaller than a given threshold value, or the number of iterations n reaches a maximum number $t_{\max}$ (lines 4-8). Finally, for each event $i \in \{1, 2, \ldots, m\}$, its state S_i is evaluated based on Eq. (14) (line 10).

3.3. Finding $\Theta^{(t+1)}$ and $\mathrm{D}^{(t+1)}$ with a fixed $S = S^{(t+1)}$

The problem in Eq. (7) with a fixed $S = S^{(t+1)}$ under $\Theta = \Theta^{(t+1)}$ at iteration t is to find the optimal estimators of unknown parameters Θ together with the most likely dependency D, so as to maximize the probability of individuals' observations A. It can be formulated as follows:

$$\begin{aligned}\Theta^{(t+1)} &\leftarrow \arg\max_{\Theta} p(\mathrm{A} \mid S; \mathrm{H}, \Theta), \\ \mathrm{D}^{(t+1)} &\leftarrow \arg\max_{\mathrm{D}} p(\mathrm{D} \mid \mathrm{A}, S; \mathrm{H}, \Theta^{(t+1)}).\end{aligned} \tag{15}$$

where $p(\mathrm{A} \mid S; \mathrm{H}, \Theta) = \sum_{\mathrm{D}} p(\mathrm{A}, \mathrm{D} \mid S; \mathrm{H}, \Theta)$. Here, $\Theta = \{R, p_s, p_d\}$, and the parameters R and p_d need to be updated when S is given.

As known from Eq. (15), the optimal estimate of Θ is related to the marginal likelihood $p(\mathrm{A} \mid S; \mathrm{H}, \Theta)$ of A. However, computing $p(\mathrm{A} \mid S; \mathrm{H}, \Theta)$ is also intractable, which considers all possible values of D. Here, an EM algorithm (called as EMD) is presented to find the MLE of the marginal likelihood $p(\mathrm{A} \mid S; \mathrm{H}, \Theta)$ of A. In EMD, A and D are the observed data and latent variables, respectively, while the parameters in Θ need to be estimated. The corresponding likelihood

Algorithm 3 EMD

Input: A, S, $\Theta^{(t+1)}$ and t_{max}.
Output: $\mathrm{D}^{(t+1)}$ and $\Theta^{(t+1)}$.
$n \leftarrow 0$ and $\Theta^{(n)} \leftarrow \Theta^{(t+1)}$.
while $||\Theta^{(n+1)} - \Theta^{(n)}|| \geq 0.01$ and $n < t_{max}$ **do**
    E-step: Compute the expectation of the log likelihood function $\mathbf{Q}(\Theta \mid \Theta^{(n)})$ of $p(\mathrm{A}, \mathrm{D} \mid S; \mathrm{H}, \Theta)$ based on Eq. (17).
    M-step: Maximize $\mathbf{Q}(\Theta \mid \Theta^{(n)})$, and then we get $\Theta^{(n+1)}$ based on Eqs. (18) and (19).
    $n \leftarrow n + 1$.
end while
$\Theta^{(t+1)} \leftarrow \Theta^{(n)}$.
Compute $\mathrm{D}^{(t+1)}$ based on Eqs. (20) and (2).

function $p(\mathrm{A}, \mathrm{D} \mid S; \mathrm{H}, \Theta)$ is computed as follows:

$$\begin{aligned} p(\mathrm{A}, \mathrm{D} \mid S; \mathrm{H}, \Theta) &= p(\mathrm{A} \mid \mathrm{D}, S; \mathrm{H}, \Theta) \cdot p(\mathrm{D}) \\ &= \prod_{j \in \mathcal{J}^0} p(\mathrm{A}_{\cdot j} \mid S; \mathrm{H}_{\cdot j}, \Theta) \cdot \prod_{j \in \mathcal{J}^1} \prod_{i \in \mathcal{E}: |\mathcal{K}_{ij}|=0} p(\mathrm{A}_{ij} \mid \mathcal{K}_{ij}, \mathrm{D}_{ij} = 0, S_i; \mathrm{H}_{\cdot j}, \Theta) \\ &\cdot \prod_{j \in \mathcal{J}^1} \prod_{a \in \{1,0\}} \prod_{i \in \mathcal{E}: |\mathcal{K}_{ij}| \neq 0, S_i = a} p(\mathrm{A}_{ij} \mid \mathcal{K}_{ij}, \mathrm{D}_{ij} = C_j^a, S_i = a; \mathrm{H}_{\cdot j}, \Theta) \cdot p(C_j^a). \end{aligned} \tag{16}$$

Next, the EMD algorithm maximizes this likelihood to estimate the parameters in Θ by iterating the two following steps: **E-step** and **M-step**.

E-step: Compute the expected log function $\mathbf{Q}(\Theta \mid \Theta^{(n)})$ of $p(\mathrm{A}, \mathrm{D} \mid S; \mathrm{H}, \Theta)$.

$$\begin{aligned} \mathbf{Q}(\Theta \mid \Theta^{(n)}) &= \mathbb{E}_{\mathbf{D} \mid \Theta^{(n)}} \Big[\ln p(\mathrm{A}, \mathbf{D} \mid S; \mathrm{H}, \Theta) \Big] \\ &= \sum_{j \in \mathcal{J}^0} \ln p(\mathrm{A}_{\cdot j} \mid S; \mathrm{H}_{\cdot j}, \Theta) + \sum_{j \in \mathcal{J}^1} \sum_{i \in \mathcal{E}: |\mathcal{K}_{ij}|=0} \ln p(\mathrm{A}_{ij} \mid \mathcal{K}_{ij}, \mathrm{D}_{ij} = 0, S_i; \mathrm{H}_{\cdot j}, \Theta) \\ &+ \sum_{j \in \mathcal{J}^1} \sum_{a \in \{1,\ 0\}} \sum_{b \in \{1,\ 0\}} p(C_j^a = b \mid \mathrm{A}_{\cdot j}, \mathcal{K}_{\cdot j}, S; \mathrm{H}_{\cdot j}, \Theta^{(n)}) \\ &\cdot \ln \sum_{i \in \mathcal{E}: |\mathcal{K}_{ij}| \neq 0, S_i = a} p(\mathrm{A}_{ij} \mid \mathcal{K}_{ij}, \mathrm{D}_{ij} = b, S_i = a; \mathrm{H}_{\cdot j}, \Theta) \cdot (p_d^a)^b \cdot (1 - p_d^a)^{(1-b)}. \end{aligned} \tag{17}$$

Here, $p(C_j^a = 1 \mid \mathrm{A}_{\cdot j}, \mathcal{K}_{\cdot j}, S; \mathrm{H}_{\cdot j}, \Theta^{(n)})$ is the conditional probability that the agent j is unfamiliar with the events that have (not) occurred under $a = 1$ ($a = 0$) conditioned on the observations of the agent and his ancestors to those events being $\mathrm{A}_{\cdot j}$ and $\mathcal{K}_{\cdot j}$, respectively, and the true state of those events being S under the current estimation of $\Theta^{(n)}$. It can be computed as follows: for each $a \in 1,\ 0$,

$$\begin{aligned} p(C_j^a = 1 \mid \mathrm{A}_{\cdot j}, \mathcal{K}_{\cdot j}, S; \mathrm{H}_{\cdot j}, \Theta^{(n)}) &= \frac{p(C_j^a = 1 \mid \mathrm{A}_{\cdot j}, \mathcal{K}_{\cdot j}, S; \mathrm{H}_{\cdot j}, \Theta^{(n)})}{\sum_{C_j^a \in \{0,\ 1\}} p(C_j^a \mid \mathrm{A}_{\cdot j}, \mathcal{K}_{\cdot j}, S; \mathrm{H}_{\cdot j}, \Theta^{(n)})} \\ &= \frac{p_d^a \cdot p(\mathrm{A}_{\cdot j} \mid \mathcal{K}_{\cdot j}, C_j^a = 1, S; \mathrm{H}_{\cdot j}, \Theta^{(n)})}{p_d^a \cdot p(\mathrm{A}_{\cdot j} \mid \mathcal{K}_{\cdot j}, C_j^a = 1, S; \mathrm{H}_{\cdot j}, \Theta^{(n)}) + (1 - p_d^a) \cdot p(\mathrm{A}_{\cdot j} \mid \mathcal{K}_{\cdot j}, C_j^a = 0, S; \mathrm{H}_{\cdot j}, \Theta^{(n)})}, \end{aligned} \tag{18}$$

where $p(\mathrm{A}_{\cdot j} \mid \mathcal{K}_{\cdot j}, C_j^a = 1, S; \mathrm{H}_{\cdot j}, \Theta^{(n)}) = \prod_i p(\mathrm{A}_{ij} \mid \mathcal{K}_{ij}, C_j^a = 1, S_i; \mathrm{H}_{\cdot j}, R_j^{(n)})$. Moreover, $p(C_j^a = 0 \mid \mathrm{A}_{\cdot j}, \mathcal{K}_{\cdot j}, S; \mathrm{H}_{\cdot j}, \Theta^{(n)}) = 1 - p(C_j^a = 1 \mid \mathrm{A}_{\cdot j}, \mathcal{K}_{\cdot j}, S; \mathrm{H}_{\cdot j}, \Theta^{(n)})$.

M-step: Maximize the expected log function $\mathbf{Q}(\Theta \mid \Theta^{(n)})$, and then get the parameters in Θ of the next iteration. We take the derivatives of $\mathbf{Q}(\Theta \mid \Theta^{(n)})$ to each parameter in Θ, and then find the optimal solutions when these derivatives

are equal to 0. And thereafter, for each agent j, and $a \in \{0,\ 1\}$, we get

$$
\begin{aligned}
(R_j^{a0})^{(n+1)} &= \frac{\sum_{i\in\mathcal{E}:|\mathcal{K}_{ij}|\neq 0,S_i=a,\mathrm{A}_{ij}=a} p(C_j^a=0 \mid \mathrm{A}_{\cdot j},\mathcal{K}_{\cdot j},S;\mathrm{H}_{\cdot j},\Theta^{(n)}) + |\{i\in\mathcal{E}:|\mathcal{K}_{ij}|=0,S_i=a,\mathrm{A}_{ij}=a\}|}{\sum_{i\in\mathcal{E}:|\mathcal{K}_{ij}|\neq 0,S_i=a} p(C_j^a=0 \mid \mathrm{A}_{\cdot j},\mathcal{K}_{\cdot j},S;\mathrm{H}_{\cdot j},\Theta^{(n)}) + |\{i\in\mathcal{E}:|\mathcal{K}_{ij}|=0,S_i=a\}|}, \\
(R_j^{a1})^{(n+1)} &= \frac{\sum_{i\in\mathcal{E}:|\mathcal{K}_{ij}|\neq 0,S_i=a,\mathrm{A}_{ij}=a} p(C_j^a=1 \mid \mathrm{A}_{\cdot j},\mathcal{K}_{\cdot j},S;\mathrm{H}_{\cdot j},\Theta^{(n)})}{\sum_{i\in\mathcal{E}:|\mathcal{K}_{ij}|\neq 0,S_i=a} p(C_j^a=1 \mid \mathrm{A}_{\cdot j},\mathcal{K}_{\cdot j},S;\mathrm{H}_{\cdot j},\Theta^{(n)})}, \\
(p_d^a)^{(n+1)} &= \frac{\sum_{j\in\mathcal{J}^1:|\mathcal{K}_{ij}|\neq 0,S_i=a,\exists i\in\mathcal{E}} p(C_j^a=1 \mid \mathrm{A}_{\cdot j},\mathcal{K}_{\cdot j},S;\mathrm{H}_{\cdot j},\Theta^{(n)})}{|\{j\in\mathcal{J}^1:|\mathcal{K}_{ij}|\neq 0,S_i=a,\exists i\in\mathcal{E}\}|}.
\end{aligned}
\tag{19}
$$

The above E-step and M-step are iteratively executed until the estimated Θ values converge. Finally, the EMD algorithm evaluates the C_j^a value of each individual $j \in \mathcal{J}^1$ to the events with $S_i = a$ as follows:

$$
(C_j^a)^{(t+1)} \leftarrow \underset{C_j^a\in\{0,\ 1\}}{\arg\max}\ p(C_j^a \mid \mathrm{A}_{\cdot j},\mathcal{K}_{\cdot j},S;\mathrm{H}_{\cdot j},\Theta^{(n+1)}). \tag{20}
$$

The framework of the EMD algorithm is shown in **Algorithm 3**. The EMD algorithm alternates between the E-step and M-step until it satisfies a convergence criterion: either the difference in norm of the estimated Θ between subsequent iterations is smaller than a given threshold value, or the number of iterations n reaches a maximum number $t_{\max}$ (lines 4-8). Finally, for each $\mathbf{D}_{ij}$, $i \in 1, 2, \ldots, m, j \in 1, 2, \ldots, q$, its value is computed based on Eqs. (20) and (2) (line 10).

4. Experimental results

In this section, experiments on synthetic data and three real-world data are given, and meanwhile a comparison of IEMTD with the benchmark algorithms for the truth discovery in social sensing is made, including MV, two-errors based expectation maximization (record as EM) [17] and DEM [31].

Benchmark algorithms. In benchmark methods, MV and EM [17] use the majority voting and expectation maximization to infer the truth of events in social sensing with independent observations, respectively. A comparison of IEMTD with MV and EM is made to demonstrate the superiority of IEMTD on the truth discovery when dependencies exist between agents' observations. DEM [31] also considers the dependency between individuals' observations, which infers the truth of events under the dependency using an EM algorithm. Note that, unlike IEMTD, DEM does not optimize the estimation of the dependency between individuals' observations. A comparison of IEMTD with DEM is made to show the effectiveness of considering the probabilistic of individuals' observations to be dependent on the truth discovery.

Metric. The truth finding accuracy, a widely used criterion to validate the performance of truth discovery algorithms in social sensing, computes the percentage of events whose inferred states are their true states. It is computed as follows:

$$
\text{Truth finding accuracy} = \sum_{i=1}^{m} \mathbf{\Delta}(S_i^*, S_i)/m,
$$

where $\mathbf{\Delta}()$ is a delta function, i.e., if $S_i^* = S_i$, $\mathbf{\Delta}(S_i^*, S_i) = 1$, and $\mathbf{\Delta}(S_i^*, S_i) = 0$ otherwise. In simulations, the value of the truth finding accuracy is in the range of [0, 1], and an algorithm with a high accuracy value has a good performance.

4.1. Experimental results on the simulation dataset

The simulation data try to model some real applications in social sensing with information flows among individuals. As that in [31], we use a forest of two level-two trees to generate the simulation data, which work as follows: First, m events are generated and each event i is set to be occurred with a prior probability p_s. Second, q individuals are divided into the root individuals and leaf individuals randomly, and the fraction of root individuals is determined by a parameter α. To model the information flows in individuals, the simulation data allow to propagate observations from the root individuals to the leaf ones. Then, each event is randomly reported by a fraction $c \in [0,\ 1]$ of individuals, and each observation A_{ij} is generated based on the reliability $R_j = \{R_j^{10}, R_j^{00}, R_j^{11}, R_j^{01}\}$ of individual j. Here, the probability for a leaf individual to give dependent observations is determined by p_d^1 and p_d^0 for the event that has occurred and not, respectively.

We test IEMTD, MV, EM and DEM on synthetic data with different parameter settings. Specifically, we take $m = 200$, $q = 100$, $p_s = 0.6$, $c = 0.4$, $\alpha = 0.4$, $R_j^{10}, R_j^{00} \sim \mathrm{Unif}(0.3, 0.8)$, $R_j^{11}, R_j^{01} \sim \mathrm{Unif}(0.3, 0.5)$, $j = 1, 2, \ldots, q$, and $p_d^1, p_d^0 \sim \mathrm{Unif}(0.3, 0.7)$, where the notation $p \sim \mathrm{Unif}(0.3, 0.8)$ means that the probability variable p follows an uniform distribution in the range of [0.3, 0.8]. For each simulation setting, 100 sets of simulation data are independently generated, and the averaged results over 100 independent trials are recorded. Next, the influence of some parameter settings is analyzed by varying one of the parameters while keeping the other parameters unchanged.

To demonstrate the superiority of IEMTD, a comparison of IEMTD with MV, EM and DEM is made on synthetic data and the comparison results are recorded in Table 2. As shown in Table 2, IEMTD has a higher truth finding accuracy than

Table 2: Truth finding accuracy of different algorithms on simulation data.

Algorithm	IEMTD	DEM	MV	EM
Truth finding accuracy	**0.7375**	0.5995	0.5288	0.5810

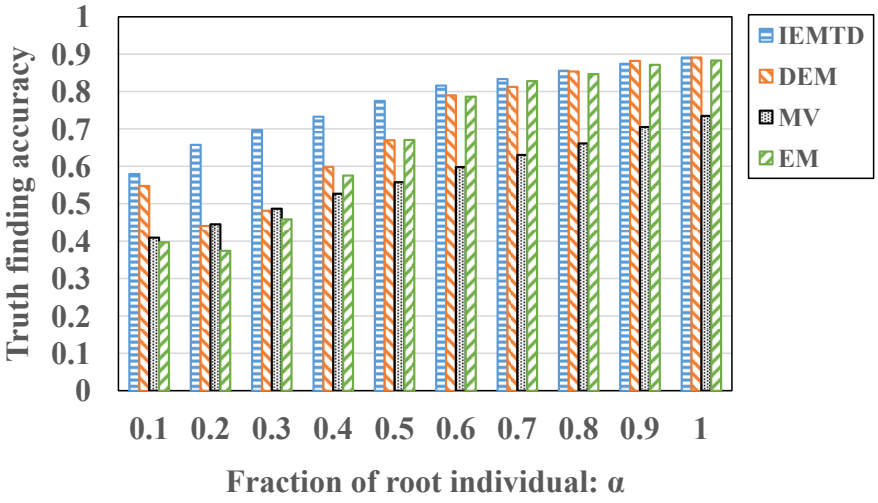
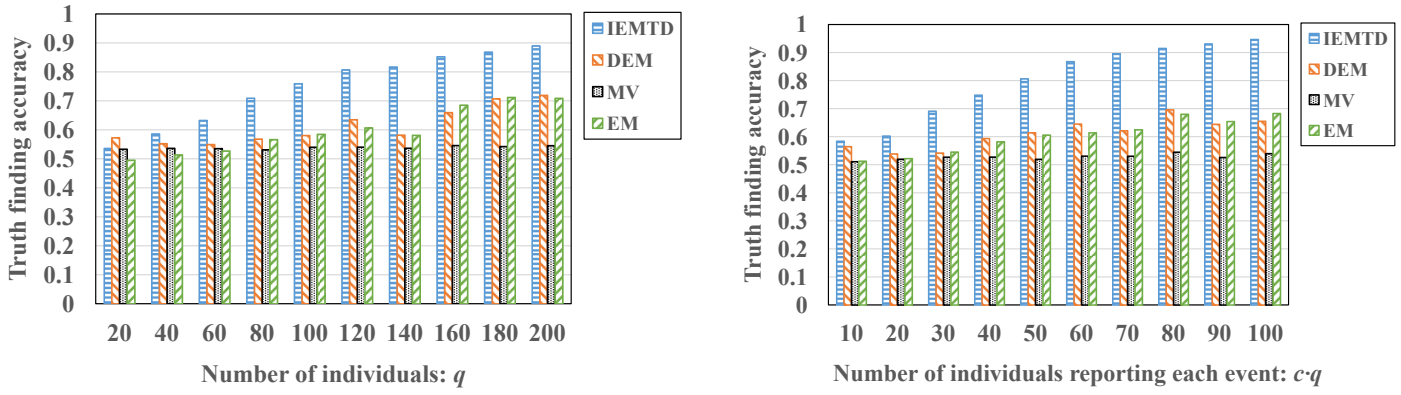


Figure 2: Truth finding accuracy VS. Fraction of root individuals.



(a) (b)

Figure 3: Truth finding accuracy VS. Number of observation information reflected by (a) the number of individuals q and (b) the number of individuals $c \cdot q$ reporting each event.

MV and EM, because it considers the dependency of individuals' observations. The results also illustrate that IEMTD has a higher truth finding accuracy than DEM due to its optimization on the dependency estimation, which considers the probabilistic of reporting an independent observation for individuals who receive information from other individuals.

The parameter α would influence the number of independent observations in the simulation data, and hence impact the inference performance of all algorithms. Fig. 2 plots the variation of the truth finding accuracy of all algorithms with α. As shown in Fig. 2, the truth finding accuracy of all algorithms varies inversely with α. This is because the number of dependent observations decreases with an increasing α value. The results also illustrate that when $\alpha \leq 0.5$, DEM, MV and EM have a low truth finding accuracy whereas IEMTD still has a high truth finding accuracy. This is because only IEMTD optimizes the estimation of the dependency between individuals' observations, which allows it to distinguish the low-reliability dependent observations, and thus reducing their influences on the truth discovery.

The parameters q and $c \cdot q$ would impact the number of observations in social sensing systems, where $c \cdot q$ denotes the number of individuals reporting an event. To investigate their influence, all algorithms are tested on simulation data with different q and $c \cdot q$ values, and the results are recorded in Fig. 3. The results demonstrate that the truth finding accuracy of EM, DEM and IEMTD increases with q and $c \cdot q$; because with the increase in the amount of observed data, the EM-based algorithms enable to reveal more useful information. The results also show the superiority of IEMTD over DEM and EM, indicating that IEMTD ia able to extract more useful information from the dependent observations due to its assumption on the probabilistic dependency of observations and its optimization on the dependency estimation.

The performance of all algorithms may also be influenced by the reliability of individuals in truth discovery systems.

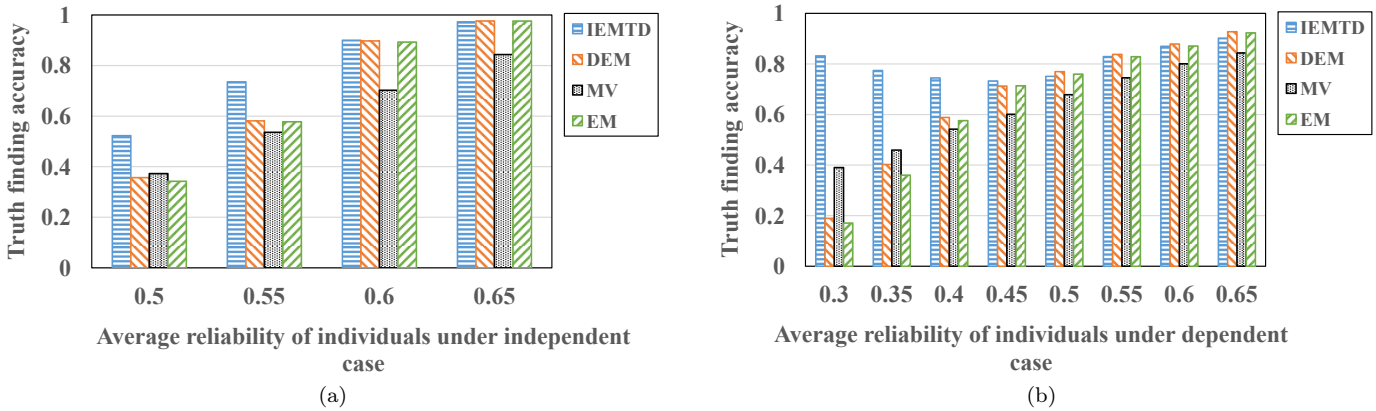


Figure 4: Truth finding accuracy VS. Average reliability of individuals. (a) the average reliability of individuals under independent case and (b) the average reliability of individuals under the dependent case.

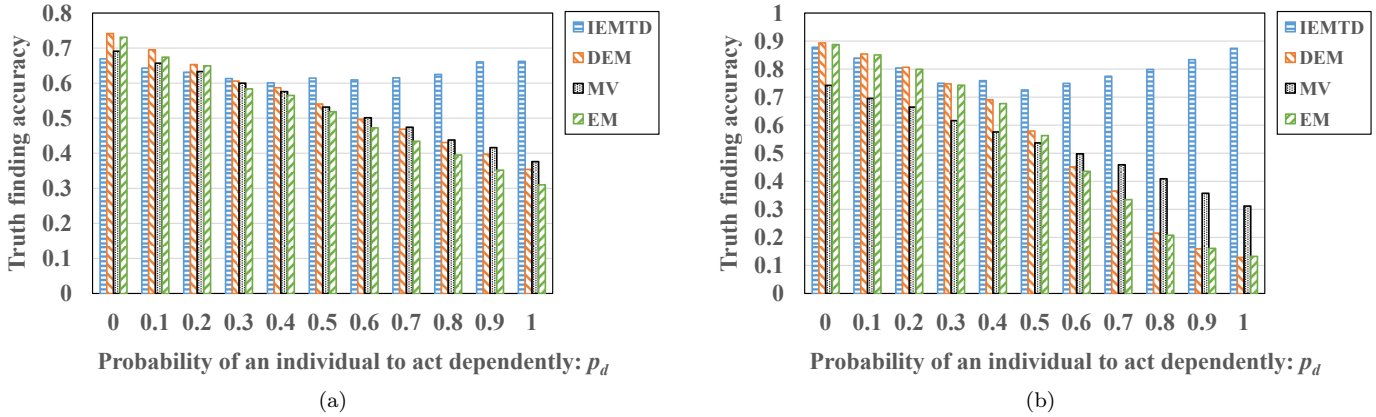


Figure 5: Truth finding accuracy VS. Probability of a leaf individual to give dependent observations with the number of individuals reporting each event (a) $c \cdot q = 20$ and (b) $c \cdot q = 40$. Here, $p_d = k$ means $p_d^1 = p_d^0 = k$.

In reality, most individuals are ordinary people with no professional knowledge about the events. Here, we set the average reliability of individuals under the independent case ranging from 0.5 to 0.65. Fig. 4a presents the performance of all algorithms varying with different reliability. From Fig. 4a, we can see that when the average reliability of individuals under the independent case is 0.5, only IEMTD's truth finding accuracy is larger than 0.5. This is because DEM, MV and EM are hard to distinguish the dependent observations which have a low reliability, thus misestimating the reliability of individuals' observations under the independent case. In this case, DEM and EM have a lower truth finding accuracy than MV, because the performance of the EM-based algorithms depends on the quality of observed data. The results also illustrate that all algorithms have a high truth finding accuracy when most observations are reliable, i.e., the average reliability is larger than 0.6. Under this case, the EM-based algorithms have a higher truth finding accuracy than MV as they can effectively infer information from reliable observation data.

Most individuals under dependent reports are between the following two extreme cases: i) spammers who simply follow their ancestors' observations and ii) wise individuals who aggregate their own observations and received observations from other individuals. Here, we set the average reliability of individuals under the dependent reports ranging from 0.3 (spammers) to 0.65 (wise individuals), and investigate its effects in Fig. 4b. From Fig. 4b, we can see that the truth finding accuracy of all algorithms increases with the increase of the reliability of individuals under the dependent case. This is expected as all the observations are more reliable on average. The results also demonstrate that when the average reliability of individuals under the dependent case is lower than 0.4, the performance of DEM and EM degrades significantly. However, in these cases, IEMTD achieves a truth finding accuracy of over 90% as it can effectively suppress the influences of spammers' unreliable observations.

The parameter p_d would influence the number of dependent observations in the system, and hence affect the truth inference accuracy of algorithms. Fig. 5 presents the truth finding accuracy of all algorithms under the settings of p_d varying from 0 to 1 (here, $p_d = k$ means $p_d^1 = p_d^0 = k$). As shown in Fig. 5, when $p_d = 0$, IEMTD does not work as well as DEM and EM because the assumptions of DEM and EM fit the data and they estimate less parameters than IEMTD. When $p_d > 0$, IEMTD performs the best in most cases as its assumptions fit the data. The results also show that the performance of DEM, MV and EM degrades sharply with the increases of p_d and $c \cdot q$ whereas that of IEMTD is consistent.

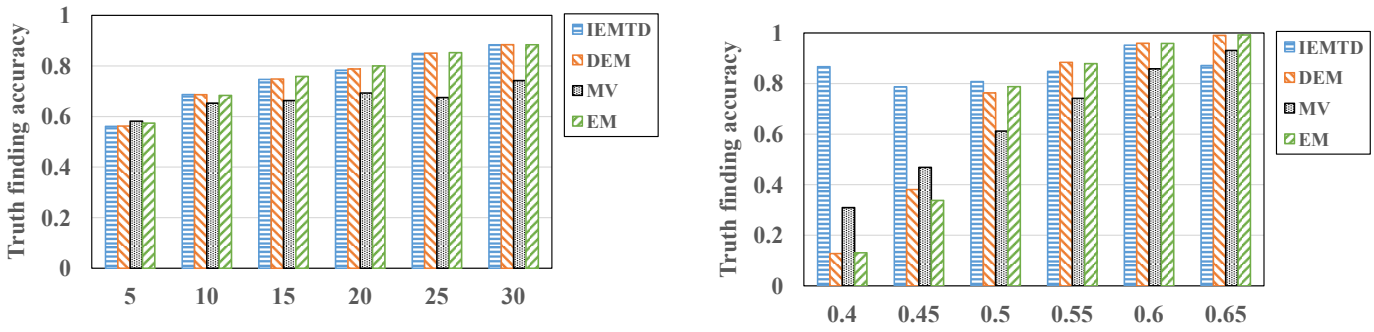


Figure 6: Truth finding accuracy VS. (a) number of individuals reporting each event with $q = 30$ at $p_d = 0$ and (b) reliability of individuals under the dependent case at $p_d = 1$.

Compared with EM, DEM needs to estimate more parameters. Therefore, with the decrease of $c \cdot q$, DEM may have a lower performance than EM, which is further validated by the results in Fig. 6a. Fig. 6a shows the variations of the performance of all algorithms with $c \cdot q$ (the number of individuals reporting each event) under $p_d = 0$ and $q = 30$. The results illustrate that when $5 \leq c \cdot q \leq 30$, EM performs the best among the EM-based truth discovery methods. In our simulation setting ($q = 100$ and $c \cdot q = 40$), DEM performs the best when $p_d = 0$. This is because the reliability parameters under different cases ($D_{ij} = 0$ and $D_{ij} = 1$) in DEM can be well estimated by a large amount of observation data and DEM can also fit the data well. Recall that the dependency in DEM is defined to be whether agents give the same observations as their ancestors. In this case, DEM can obtain a better performance than EM, which was validated by [31].

Fig. 6b records the variations of the performance of all algorithms with the average reliability of individuals under the dependent case at $p_d = 1$. As shown in Fig. 6b, when this average reliability is larger than 0.50, IEMTD obtains a lower truth finding accuracy than DEM because the DEM fits the data and it estimates less parameters than IEMTD. When the average reliability is smaller than 0.45, IEMTD obtains a higher truth finding accuracy than DEM. This is because the difference of the reliability under between the independent case and dependent case makes IEMTD's estimated reliability different from that of DEM. Moreover, the EM-algorithms cannot guarantee convergence to a global optimum, but they ensure that the marginal likelihood of the observed data is non-decreasing over iterations.

To further verify the availability of IEMTD, we test all algorithms on a special simulation with the same average reliability 0.5 for all observations. In this case, each claim can be considered to be independent or dependent. IEMTD may have a lower truth finding accuracy than the benchmark algorithms, because the optimization of the estimation of dependency becomes useless and consumes some computational resources from a finite resource. This is further demonstrated by the corresponding experimental results, i.e., the truth finding accuracy of IEMTD, DEM, MV and EM is 0.6458, 0.6507, 0.6130 and 0.6535, respectively. Note that, IEMTD has an acceptable truth finding accuracy, which is higher than MV and close to DEM and EM.

4.2. Experimental results on the real-world dataset

We test all algorithms on three Twitter datasets collected by [31] in 2015 with different keywords, i.e., Ukraine, Kirkuk and Pairs Attack. A brief description of these datasets is given in Table 3, and the details are shown as follows:

Ukraine: On March 14th 2015, the Russian President Vladimir V. Putin has not been seen in public for nine days. During the nine days, Putin's office has cancelled a series of meetings and visits, e.g., the meeting with the Federal Security Service and the scheduled visit to Kazakhstan, the Kazakh government official told Reuters that the Russian President Vladimir V. Putin was ill, an anonymous individual in the elite Moscow Central Clinical Hospital has claimed that Putin suffered ischaemic stroke, etc. These news provoked a series of events (to facilitate the discussions, hereafter we refer to all the related news reported on the studied social media, true or not, as "events") in news and social media, e.g., the Russia president was dead, he suffered from heart attack, he was ill with cancer, he was fine and attended the meeting with the president of the Republic of Karelia, etc. Most of these events were not true, yet they were followed by a set of individuals in Twitters. This dataset can be used to test the performance of the benchmark algorithms on discovering the truth of events when individual's observations are dependent. In this dataset, if an individual is unfamiliar (familiar) with the event "Putin is fine", then his observations will be (not be) influenced by other individual's observations.

Kirkuk: On March 9th 2015, ISIS around Kirkuk in northern Iraq suffered from the attack from Kurdish forces. On March 10th 2015, a set of events about the attack, which were not totally true, were discussed and followed on social

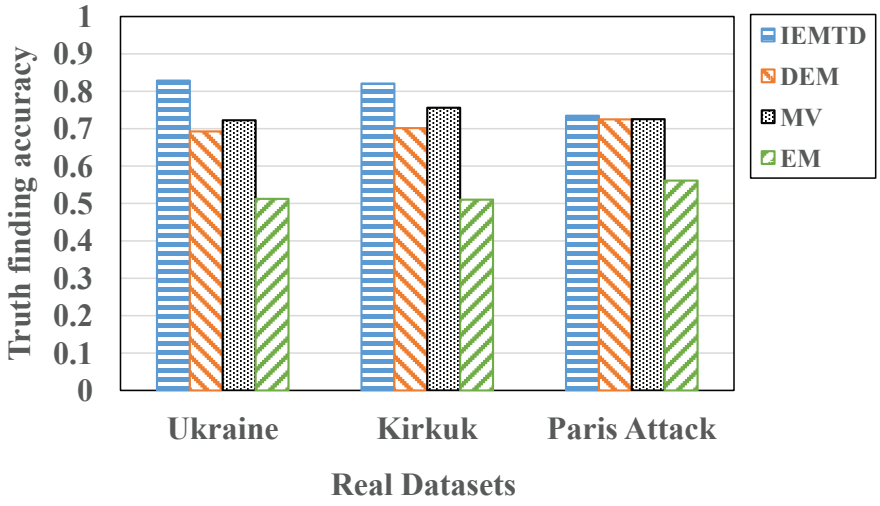


Figure 7: Truth finding accuracy of each algorithm on real-world datasets.

Table 3: Brief information of the tested real-world Twitter datasets. Here, the notation ♯ represents the number of corresponding entities, e.g., ♯ Event denotes the number of events.

Dataset	Start-End Time	Evaluation Day	♯ Event	♯ Individual	♯ Observation	♯ Information flow
Ukraine	Feb 20 2015- Mar 31 2015	Mar 31 2015	393	180	1211	142
Kirkuk	Jan 31 2015- Apr 02 2015	Mar 10 2015	328	190	996	289
Paris Attack	Nov 14 2015- Mar 31 2015	Nov 24 2015	1304	336	3924	255

media, e.g., the attack was cooperated with the air strikes from an U.S.-led coalition, 20 Shi'ite militiamen were killed, etc.

Paris Attack: On November 13th 2015, a series of premeditated terrorist attacks occurred in Paris, France. In 40 min, three explosions occurred successively in the public areas of Paris. These attacks killed more than 130 people and injured 368 more. After these attacks, many events were discussed and followed on social media. We extract those events on November 21 to test the performance of the benchmark algorithms.

For actual evaluation, similar to the data processing in [31], we collected the tweets with follower-followee relationships and ranked in top-5000 according to their computed truth probabilities. The real states of collected events were manually graded as "True" and "False" based on the background research of graders on each tweet.

The truth finding accuracy of each algorithm is shown in Fig. 7. The results illustrate that IEMTD outperforms the benchmark algorithms. On average, IEMTD has improved at least 14.68%, 8.57% and 1.28% of the truth finding accuract of the benchmark algorithms for the real-world Ukraine, Kirkuk and Paris Attack data, respectively. The results also show that DEM has a higher truth finding accuracy than EM, but it has a lower accuracy than IEMTD. This is to be expected as DEM considers the dependency between individual's observations whereas it does not optimize the estimation of dependency. This phenomenon further demonstrates the effectiveness of considering the probabilistic dependency of observations and optimizing the dependency in IEMTD.

As shown in Fig. 7, for real data, MV has a higher truth finding accuracy than EM and DEM. This is because the

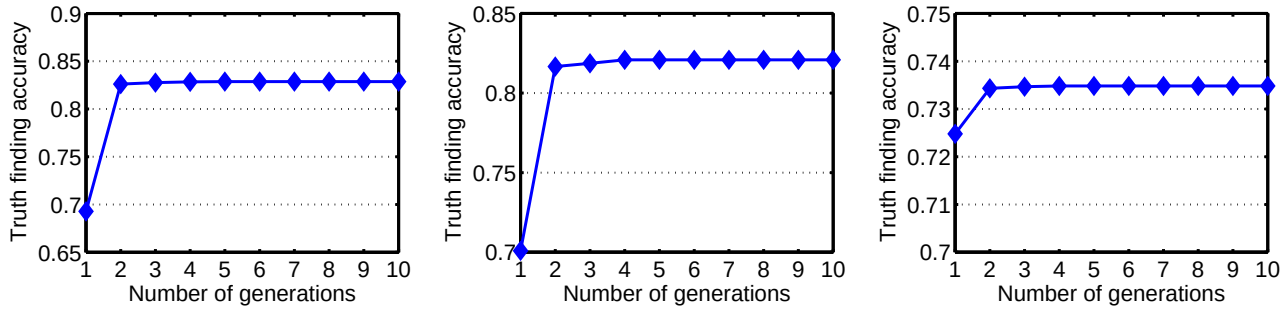


Figure 8: Convergence performance of IEMTD on real-world (a) Ukraine data, (b) Kiruku data and (c) Paris Attack data.

EM-based algorithms have many unknown parameters to be estimated from the observed data, hence their performance strongly depends on the amount of observation data. This was demonstrated by the results in Fig. 3. As shown in Fig. 3, the EM-based algorithms have a low truth finding accuracy similar to MV when $c \cdot q = 10$. The tested Twitter data have a few observations ($c \cdot q < 10$) for each individual. Moreover, there exist information flows in individuals. These factors make the EM-based algorithms have a lower truth finding accuracy than MV.

Fig. 8 presents the truth finding accuracy obtained by IEMTD after each iteration on the three real-world datasets. In the first iteration, IEMTD infers the truth of events under an initial D generated by MV. In this case, its truth finding accuracy is about 0.7 for the three real datasets similar to that of DEM. After that, IEMTD estimates the truth of events under an optimal D found by its former iteration. As shown in Fig. 8, the truth finding accuracy of IEMTD is improved obviously at the second iteration under the optimal D returned by the function **EMD** of IEMTD at the first iteration, demonstrating the effectiveness of the optimization of the estimation of dependency. The results also illustrate that IEMTD can converge within ten iterations for the three real datasets (IEMTD produces no further improvements for the truth finding accuracy when $t_{max} > 5$).

5. Conclusions

In this paper, we have developed a truth discovery approach to infer the true state of events from observations provided by unreliable individuals, who may be influenced by other individuals. We have proposed an iterative expectation maximization algorithm to infer the truth of the observed events and the reliability of individuals. The algorithm takes into account the interplay between the reliability of individuals and the dependency of individuals' observations. Extensive experiments using both simulations and real data have demonstrated that our method outperforms three classical truth discovery methods in the literature, when dependencies exist between individuals' observations. In future work, we will investigate the historical data for truth discovery to infer the dependence relationships, and the correlation (e.g., social connection of individuals and the coupling of events) between individuals and events.

Acknowledgements

The authors would like to thank the authors in [31] for their providing Twitter datasets. This work was supported in part by the Singapore Ministry of Education Academic Research Fund Tier 2 grant MOE2014-T2-1-028.

References

[1] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Göker, I. Kompatsiaris, A. Jaimes, Sensing trending topics in twitter, IEEE Transactions on Multimedia 15 (6) (2013) 1268–1282.

[2] S. Chatterjee, A. Mukhopadhyay, M. Bhattacharyya, Dependent judgment analysis: A markov chain based approach for aggregating crowdsourced opinions, Information Sciences 396 (2017) 83–96.

[3] M. Christopher, Pattern recognition and machine learning, Springer Press, Springer-Verlag New York, 2016.

[4] X. L. Dong, L. Berti-Equille, D. Srivastava, Integrating conflicting data: the role of source dependence, Proceedings of the VLDB Endowment 2 (1) (2009) 550–561.

[5] X. L. Dong, L. Berti-Equille, D. Srivastava, Truth discovery and copying detection in a dynamic world, Proceedings of the VLDB Endowment 2 (1) (2009) 562–573.

[6] A. Godoy-Lorite, R. Guimerà, C. Moore, M. Sales-Pardo, Accurate and scalable social recommendation using mixed-membership stochastic block models, Proceedings of the National Academy of Sciences 113 (50) (2016) 14207–14212.

[7] C. Huang, D. Wang, Topic-aware social sensing with arbitrary source dependency graphs, in: Proceedings of the 15th IEEE International Conference on Information Processing in Sensor Networks, Vienna, Austria, 2016.

[8] C. Huang, D. Wang, N. Chawla, Towards time-sensitive truth discovery in social sensing applications, in: Proceedings of the 12th IEEE International Conference on Mobile Ad Hoc and Sensor Systems, Dallas, USA, 2015.

[9] C. Huang, D. Wang, N. Chawla, Scalable uncertainty-aware truth discovery in big data social sensing applications for cyber-physical systems, IEEE Transactions on Big Data 1 (2017) 1–14.

[10] V. Krishnamurthy, H. V. Poor, A tutorial on interactive sensing in social networks, IEEE Transactions on Computational Social Systems 1 (1) (2014) 3–21.

[11] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, A. T. Campbell, A survey of mobile phone sensing, IEEE Communications magazine 48 (9) (2010) 140–150.

[12] J. Lee, D. Lee, S.-w. Hwang, Crowdk: Answering top-k queries with crowdsourcing, Information Sciences 399 (2017) 98–120.

[13] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, J. Han, A confidence-aware approach for truth discovery on long-tail data, Proceedings of the VLDB Endowment 8 (4) (2014) 425–436.

[14] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, J. Han, Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation, in: Proceedings of the ACM SIGMOD international conference on Management of data, New York, USA, 2014.

[15] Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, J. Han, On the discovery of evolving truth, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, 2015.

[16] Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, J. Han, Conflicts to harmony: A framework for resolving conflicts in heterogeneous data by truth discovery, IEEE Transactions on Knowledge and Data Engineering 28 (8) (2016) 1986–1999.

[17] Q. Liu, J. Peng, A. T. Ihler, Variational inference for crowdsourcing, in: Advances in neural information processing systems, Lake Tahoe, Nevada, USA, 2012.

[18] C. Meng, W. Jiang, Y. Li, J. Gao, L. Su, H. Ding, Y. Cheng, Truth discovery on crowd sensing of correlated entities, in: Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems, Seoul, Republic of Korea, 2015.

[19] P. Michelucci, J. L. Dickinson, The power of crowds, Science 351 (6268) (2016) 32–33.

[20] R. W. Ouyang, L. M. Kaplan, A. Toniolo, M. Srivastava, T. J. Norman, Parallel and streaming truth discovery in large-scale quantitative crowdsourcing, IEEE Transactions on Parallel and Distributed Systems 27 (10) (2016) 2984–2997.

[21] R. W. Ouyang, M. Srivastava, A. Toniolo, T. J. Norman, Truth discovery in crowdsourced detection of spatial events, IEEE Transactions on Knowledge and Data Engineering 28 (4) (2016) 1047–1060.

[22] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: real-time event detection by social sensors, in: Proceedings of the 19th ACM international conference on World wide web, New York, USA, 2010.

[23] H. Shin, J. Lee, Impact and degree of user sociability in social media, Information Sciences 196 (2012) 28–46.

[24] E. Soltanmohammadi, M. Naraghi-Pour, Fast detection of malicious behavior in cooperative spectrum sensing, IEEE Journal on Selected Areas in Communications 32 (3) (2014) 377–386.

[25] B. Suh, L. Hong, P. Pirolli, E. H. Chi, Want to be retweeted? large scale analytics on factors impacting retweet in twitter network, in: Proceedings of the 2nd IEEE International Conference on Social Computing, Minneapolis, MN, USA, 2010.

[26] 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 the 13th IEEE International Conference on Information Processing in Sensor Networks, Berlin, Germany, 2014.

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

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

[29] D. Wang, L. Kaplan, H. Le, T. Abdelzaher, On truth discovery in social sensing: A maximum likelihood estimation approach, in: Proceedings of the 11th IEEE International Conference on Information Processing in Sensor Networks, Beijing, China, 2012.

[30] S. Wang, L. Su, S. Li, S. Hu, T. e. a. Amin, Scalable social sensing of interdependent phenomena, in: Proceedings of the 14th IEEE International Conference on Information Processing in Sensor Networks, Seattle, USA, 2015.

[31] S. Yao, S. Hu, S. Li, Y. Zhao, L. Su, L. Kaplan, A. Yener, T. Abdelzaher, On source dependency models for reliable social sensing: Algorithms and fundamental error bounds, in: Proceedings of the 36th IEEE International Conference on Distributed Computing Systems, Nara, Japan, 2016.

[32] X. Yin, J. Han, S. Y. Philip, Truth discovery with multiple conflicting information providers on the web, IEEE Transactions on Knowledge and Data Engineering 20 (6) (2008) 796–808.

[33] J. Zhang, V. S. Sheng, J. Wu, X. Wu, Multi-class ground truth inference in crowdsourcing with clustering, IEEE Transactions on Knowledge and Data Engineering 28 (4) (2016) 1080–1085.

[34] B. Zhao, B. I. Rubinstein, J. Gemmell, J. Han, A bayesian approach to discovering truth from conflicting sources for data integration, Proceedings of the VLDB Endowment 5 (6) (2012) 550–561.