Multi-truth Discovery with Correlations of Candidates in Crowdsourcing Systems

Hongyu Huang, Guijun Fan, Yantao Li(✉), and Nankun Mu

College of Computer Science, Chongqing University, Chongqing 400044, China
yantaoli@cqu.edu.cn

Abstract. In the past decade, crowdsourcing has emerged as a popular internet-based collaborative computing paradigm. In crowdsourcing systems, requesters can ask users ('sources') for true values ('truths') of objects or events. Generally, an object may have multiple truths and sources could provide inconsistent or even conflicting answers ('candidates') about the object. For this scenario, the multi-truth discovery is a promising technique to deal with various candidates provided by different sources. However, most of the existing multi-truth discovery methods ignore the correlation between candidates so that the inferred truth could be different from the ground truth. In order to solve this problem, we propose MTD-CC, a Multi-Truth Discovery with Candidate Correlations. Specifically, we first design a metric of potential function to measure the correlation between each pair of candidates based on sources' votes and reliabilities. Then, we construct a Markov Random Field (MRF) to represent these correlations. Next, we transform the MRF into a directed graph and cut it based on the Min-cut theorem to infer which candidates are truths. Last, we evaluate the proposed method on both real and synthetic datasets and experimental results demonstrate that the accuracy of MTD-CC outperforms existing solutions.

Keywords: Crowdsourcing systems · Multi-truth discovery · Candidate correlations · Markov random field

1 Introduction

Crowdsourcing, as a popular internet-based collaborative computing paradigm, can utilize the intelligence of internet users to solve some problems that are difficult for machines. At present, there are many successful crowdsourcing platforms, such as Wikipedia, Zhihu and Baidu Zhidao. In crowdsourcing systems, people can raise a question to the public and then obtain its answers according to the responses from different users which are usually called 'data sources' or 'sources'. For examples, a patient may ask other people about his symptoms through crowdsourcing platforms before going to see a doctor, or a traveller may enquire a hotel condition before setting off. After receiving responses from

H. Gao and X. Wang (Eds.): CollaborateCom 2021, LNICST 407, pp. 18–32, 2021.
https://doi.org/10.1007/978-3-030-92638-0_2

sources, the requester can extract some 'candidates' to further infer the true answers or 'truths' of the question. Due to the diversity of sources, however, the extracted candidates are often inconsistent or even conflicting.

A simple but widely used method for aggregating different data is majority voting, where the most frequently voted candidates are regarded as truths. However, this approach ignores the quality of candidates from different sources, which leads to low accuracy of data aggregation. To address this problem, the truth discovery, which infers the truth more efficiently and accurately, has emerged as a hot topic and been applied in many domains [11], such as healthcare [14], knowledge base [1], and crowd sensing [2]. The intuition of truth discovery is that if a source provides accurate candidates frequently, it will be rewarded with high reliability. On the other side, if a candidate is supported by reliable sources, it will be highly inferred as a truth.

Most of the existing truth discovery methods are applied to single-truth scenarios. In real applications, however, it is common that an object has one or more truths, such as the side effects of a medicine, and the authors of a book. There are a few works focusing on the multi-truth discovery problem. Zhao *et al.* first observe that each source may have two types of errors, i.e., false positive and false negative. Therefore, they model two different aspects of source reliability to improve the accuracy of truth discovery [21]. Lin *et al.* propose an integrated Bayesian approach incorporating sources' domain expertise and confidence scores of value sets to find possible truths [12]. Fang *et al.* adopt a graph model that incorporates source relations, object popularity, loose mutual exclusion, and long-tail phenomenon to estimate the truth [4]. The aforementioned approaches assume that candidates are independent, but they might be correlated in real applications and their correlations could be used to further improve the inference accuracy.

Inspired by this observation, we attempt to discover multi-truth from different candidates utilizing their correlations. Therefore, we propose MTD-CC, a Multi-Truth Discovery with Candidate Correlations in this paper. Specifically, we first design a metric of potential function to quantify the correlation of candidates, which satisfies submodularity, i.e., if both candidates are supported or not supported by a source, their correlations will be increased. Then, we utilize an undirected graph model, i.e., the Markov Random Field (MRF), to construct the candidate correlation network. Next, we assign each edge of the MRF with a direction by two additional vertexes, i.e., a source vertex and a sink vertex, and the related directed edges are also added into the network. Last, we cut the network into two partitions based on the Min-cut theorem. The candidates who are connected with the sink vertex are regarded as truths. We finally evaluate MTD-CC on both real and synthetic datasets and the experimental results show that MTD-CC is notably more accurate than baseline methods.

The rest of the paper is organized as follows. Section 2 reviews the state-of-the-art. Section 3 introduces the key concepts of our problem. Section 4 presents MTD-CC in detail. Performance evaluation on real and synthetic datasets is conducted in Sect. 5. Last, we conclude our work in Sect. 6.

2 Related Work

There are state-of-the-art approaches proposed to solving the truth discovery problem and multi-truth discovery problem respectively.

2.1 Truth Discovery

There have been various methods to solve the truth discovery problem. Truth finder is the first work to formulate the truth discovery problem, and it utilizes a Bayesian-based method to compute the source reliability and object truth iteratively [19]. Li *et al.* consider the long-tail phenomenon on source coverage for objects, and confidence interval is used to represent the source reliability [9]. Liu *et al.* employ a probabilistic model to characterize the reliability of sources and propose an user recruitment algorithm, aiming to guarantee the accuracy of results in truth discovery [13]. Zheng *et al.* model the reliability of sources based on the difficulty of objects dynamically to estimate the true value [22].

However, these methods ignore correlations among sources, objects, or answers, which are helpful to improve the performance.

2.2 Truth Discovery with Different Correlations

Different kinds of correlations are often considered in the process of truth discovery. In [17], the authors propose a probabilistic graph model considering object correlations and the correlations among objects can be measured by spatial distance. Li *et al.* study the existence of clusters in sources, and divide clusters based on sources' relative locations [3]. In [18], CTD captures correlations of attributes through life experience. For example, departure time of the flight must precede its landing time. The authors in [10] measures the similarity of answers by capturing key words in different objects.

Nevertheless, the above works ignore correlations between candidates. Therefore, even if we can directly use these methods into multi-truth discovery scenario, the performance is not competitive.

2.3 Multi-truth Discovery

There are a few approaches addressing the multi-truth discovery problem. Wang *et al.* adopt a probabilistic graph model to infer true values and the source reliability and improve the performance by assuming a prior distribution of the truths [16]. In [12,15], the authors propose integrated Bayesian approaches to jointly reason about truths and the source reliability. Wang *et al.* consider the source confidence and finer-grained copy detection [15]. Lin *et al.* take domain expertise of sources into account and assign reasonable scores to value sets [12]. Fang *et al.* utilize a Markov chain to build the relations among sources and incorporate four important implications to achieve better accuracy [4].

We notice that the above approaches assume that candidates are independent so they do not measure the correlation between these candidates. Different from the aforementioned works, we consider the correlation of candidates to further improve the accuracy of multi-truth discovery.

3 Preliminaries

In this section, we first describe the key concepts of the source, candidates, and truths, and then utilize an illustrative example to give an insight understanding of these concepts.

Source. In this work, we regard the source as an entity which provides the information of an object or event. In a crowd-sensing scenario, the source could be a person who takes a sensing task and submits data to the server. In a data analysis scenario, the source could be a website which provides the authors of a book or the symptoms of a disease. Formally, we use $S = \{s_1, s_2, ..., s_N\}$ to represent the set of sources, where N is the total number of sources. Considering the trustworthiness of each source, we use $R = \{r_1, r_2, ..., r_N\}$ to denote sources' reliabilities. A source with higher reliability indicates that the corresponding data are more reliable.

Candidates. Given a certain object, different sources may provide various information about it. We assume that, before truth discovery, the information has been preprocessed by some techniques and thus we can obtain a set of key words or values about the object. For each key word, we define a binary random variable, i.e., candidate, to represent whether it is the true value of the object. Formally, we define $c_i \in \{0, 1\}$ as the i^{th} candidate. Let $c_i = 1$ or simply as c_i^1 indicate that the i^{th} key word related to c_i is a true value, where c_i^0 indicates that it is a false value. Let $C = \{c_1, c_2, ..., c_M\}$ denote the set of candidates, where M is the total number of candidates.

Truths. We assume that each object has one or more true values, i.e., ground truths, which are unknown to us. Based on the data submitted by sources, the goal is to compute the inferred truth, i.e., to infer each c_i is c_i^1 or c_i^0.

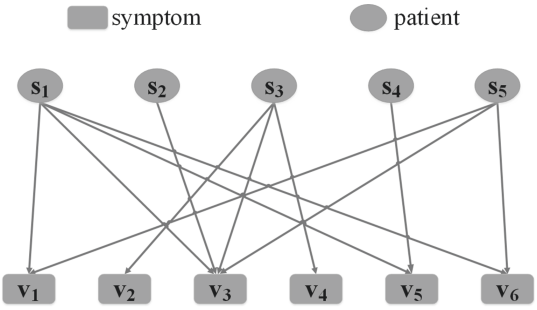


Fig. 1. Five patients report their symptoms about the side effects of a new medicine. Each directed edge represents that a patient has a certain symptom.

Example. We take an example, as shown in Fig. 1 and Table 1, to further illustrate the aforementioned concepts. Supposing we attempt to know the side effects of a new medicine from five patients ($s_1 \sim s_5$), we need to require them to provide their symptoms ($v_1 \sim v_6$) to us. Each patient supports part of the symptoms, e.g., patient s_1 reports symptoms $\{v_1, v_3, v_5, v_6\}$. We assign each symptom to a candidate and convert each patient's report as a vector of candidates, i.e., $<c_1, c_2, c_3, c_4, c_5, c_6>$, where c_i belongs to c_i^0 or c_i^1. Then, we can list each patient's report in Table 1.

Table 1. The binary representation of each patient's report.

Patients	*Candidates*					
	c_1	c_2	c_3	c_4	c_5	c_6
s_1	1	0	1	0	1	1
s_2	0	0	1	0	0	0
s_3	0	1	1	1	0	0
s_4	0	0	0	0	1	0
s_5	1	0	1	0	0	1

4 The Design of MTD-CC

In this section, we introduce the details of our proposed multi-truth discovery method, MTD-CC. Our method is based on a probabilistic graph model, i.e., the MRF, where each vertex represents a candidate taking value in $\{0, 1\}$. The edges stand for correlations between pairs of candidates. We first design a metric of potential function to measure these correlations. Since we regard each candidate as a random binary variable, the goal is to find an assignment for each candidate so that their joint probability is optimal. To achieve this goal, we convert the MRF into an undirected graph and then separate the graph into two partitions based on the Min-cut theorem, where one partition is assigned with 1 and the other with 0. Last, we regard those candidates with value 1 as truths.

4.1 Metric for Candidate Correlations

To measure the correlation between each pair of candidates, we design the potential function that takes the sources' reliabilities into account. Given two candidates c_j and c_k ($j \neq k$), there are four cases of their combination. Accordingly, we can separate sources into four sets. The set $S_{(x,y)}$, where $x, y \in \{0, 1\}$, has those sources who support x and y for c_j and c_k, respectively. Therefore, there could be four specific functions, and each of them corresponds to one case. Table 2 describes how to compute the potential function for each case.

The intuition of the potential function is that if a source votes two candidates as 1 or 0 simultaneously, we should add its reliability into the correlation. On

the contrary, if a source votes two candidates differently, an intermediate weight is reasonable. Moreover, the case with double 1s has more weights than that with double 0s because two positive votes show more significant correlation than negative votes.

Table 2. The potential function $\psi(c_j, c_k)$ of two candidates.

$\psi(c_j^0, c_k^0)$	$\sum_{i \in S_{(0,0)}} (1 - r_i)$
$\psi(c_j^0, c_k^1)$	$\sum_{i \in S_{(0,1)}} 0.5$
$\psi(c_j^1, c_k^0)$	$\sum_{i \in S_{(1,0)}} 0.5$
$\psi(c_j^1, c_k^1)$	$\sum_{i \in S_{(1,1)}} r_i$

4.2 Construction of MRF

MRF is an undirected graphical model, where each vertex depicts a random variable and the edge between them indicates their relation. In our problem, each vertex is a candidate. We define a parameter δ to determine whether two vertexes have an edge. Let $\delta_{j,k} = \psi(c_j^0, c_k^0) + \psi(c_j^1, c_k^1) - \psi(c_j^0, c_k^1) - \psi(c_j^1, c_k^0)$. Two vertexes c_j and c_k are connected with an edge if and only if $\delta_{j,k} > 0$. For two connected vertexes, Table 2 is also used to be the edge potential function of MRF.

Take c_1 and c_3 in Table 1 as an example. We can see that s_1 and s_5 vote two 1s, s_4 votes two 0s, and s_2 and s_3 vote a 1 and a 0. We assume that all sources' reliabilities are equal to 0.6, that is, $\delta_{1,3} = 0.6 + 0.6 + 0.4 - 0.5 - 0.5 = 0.6 > 0$. So we connect c_1 and c_3 with an edge. In this way, using the data in Table 1, we build an MRF which is shown in Fig. 2.

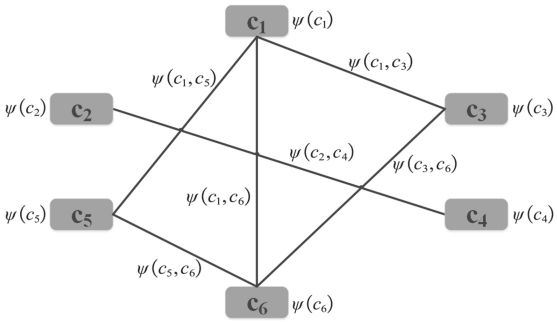


Fig. 2. MRF on a new medicine's side effects based on values provided by patients.

Next, we define the vertex potential function to quantify the extent of how a candidate is supported or opposed by sources. Similar to the edge potential

function, we use a table to define the vertex potential function, as shown in Table 3. To compute the potential function for a candidate, e.g., c_j, we first separate sources into two sets by their votes. The sets $S_{(0)}$ and $S_{(1)}$ have those sources whose votes on c_j are 0 and 1, respectively. Then, we use the sum of the square roots of their reliabilities, $\sum_{i \in S_{(0)}} \sqrt{1 - r_i}$ or $\sum_{i \in S_{(1)}} \sqrt{r_i}$, as the $\psi(c_j)$.

Table 3. The potential function $\psi(c_j)$ of a candidate.

$\psi(c_j^0)$	$\sum_{i \in S_{(0)}} \sqrt{1 - r_i}$
$\psi(c_j^1)$	$\sum_{i \in S_{(1)}} \sqrt{r_i}$

We can see that both potential functions of edges and vertexes depend on the sources' reliabilites. Initially, we assume that each source has the same reliability. When obtaining the inferred truths, we can update the sources' reliabilites by Eq. (1), and use them to iteratively infer truths.

$$r_i = g(\frac{\sum_{j \in M} I(c_{i,j}, t_j)}{M}), \tag{1}$$

where $g(\cdot)$ is a monotonically increasing function, and $I(\cdot, \cdot)$ is an indicator function. Parameter $c_{i,j}$ is the binary vote from the i^{th} source for the j^{th} candidate, and t_j represents the inferred truth for the j^{th} candidate. If the vote $c_{i,j}$ agrees with the inferred truth t_j, then $I(c_{i,j}, t_j) = 1$. Otherwise, $I(c_{i,j}, t_j) = 0$. Last, parameter M represents the number of candidates.

4.3 Transformation of MRF

Since we regard candidates as random binary variables, there are 2^M cases of their combinations in total. For each case, we can compute its probability as:

$$P(c_1, ..., c_M) = \frac{1}{Z} \prod_{c_j \in U} \psi(c_j) \prod_{(c_j, c_k) \in E} \psi(c_j, c_k), \tag{2}$$

where 'Z' is a normalized factor:

$$Z = \sum_{c_1, ..., c_M} \prod_{c_j \in U} \psi(c_j) \prod_{(c_j, c_k) \in E} \psi(c_j, c_k). \tag{3}$$

The intuition of our inference approach is to find the most possible truths by maximizing the joint probability. Then we utilize Maximum A Posteriori (MAP) estimation to infer truths T:

$$T = arg \max_{c_1, ..., c_M} P(c_1, ..., c_M) \tag{4}$$

The MAP problem is usually transformed into an energy minimization problem [7], which has two advantages. First, it avoids the numerical problems associated with multiplying a lot of small numbers. More importantly, the multiplication can be transformed into a linear computation. The energy function f is

computed as the negative likelihood of the joint probability:

$$f(c_1, ..., c_M) = -\ln(arg \max_{c_1,...,c_M} P(c_1, ..., c_M)) \propto arg \min_{c_1,...,c_M} \sum_{c_j \in U} \theta(c_j) + \sum_{(c_j,c_k) \in E} \theta(c_j, c_k) \quad (5)$$

where $\theta(c_j) = -\ln \psi(c_j)$ and $\theta(c_j, c_k) = -\ln \psi(c_j, c_k)$. Obviously, a brute-force algorithm that checks all possible joint assignments is inefficient, and thus we need to find a more efficient inference algorithm.

Our method is inspired by the Min-cut theorem, which is widely used to solve the optimization problem. For energy functions that satisfy graph-representable [8], the Min-cut can guarantee that the optimal solution will be obtained in polynomial time regardless of the number of vertexes and edges. Assume that a directed graph $\hat{G}$ with nonnegative edge weights has two terminals, i.e., a source vertex $\hat{s}$ and a sink vertex $\hat{t}$. When we divide vertexes of the graph into two disjoint subsets $\hat{S}$ and $\hat{T}$, where $\hat{s} \in \hat{S}$ and $\hat{t} \in \hat{T}$, the Min-cut requires that the sum of the edges in the cut set is minimum.

In order to utilize the Min-cut to solve the energy minimization problem, we need to construct a directed graph $\hat{G}(\hat{U}, \hat{E})$, where $\hat{U} = \{c_1, ..., c_M, \hat{s}, \hat{t}\}$ and $\hat{E}$ is transformed from the edges in the MRF. The process includes two steps. First, we design a method to transform each edge in the MRF into a directed one and assign weights. The second step is to add new directed edges with weights between the original vertexes and the added terminal vertexes. After cutting the graph $\hat{G}$ into two disjoint subsets $\hat{S}$ and $\hat{T}$, we assign 0 to vertexes which connect with $\hat{s}$ and 1 to vertexes which connect with $\hat{t}$, respectively.

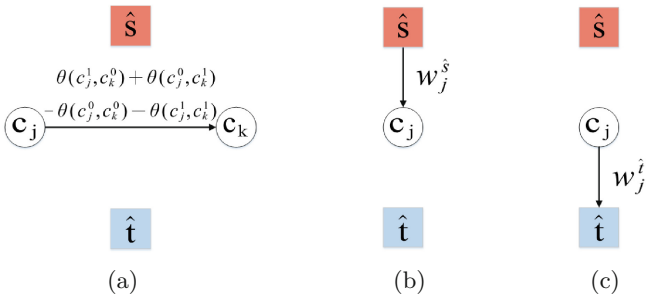


Fig. 3. Transform an MRF into a directed graph. (a) Transform an undirected edge of the MRF into a directed edge. (b) Connect the source vertex $\hat{s}$ with c_j. (c) Connect c_j with the sink vertex $\hat{t}$.

Now we introduce these two steps in detail. First, we consider two candidates c_j and c_k, where $j < k$, connected by an edge in the MRF. We transform the undirected edge to a directed one from c_j to c_k, i.e., (c_j, c_k). The weight of this edge is equal to $\theta(c_j^1, c_k^0) + \theta(c_j^0, c_k^1) - \theta(c_j^0, c_k^0) - \theta(c_j^1, c_k^1)$, as shown in Fig. 3(a).

Second, we investigate how to connect original vertexes with two terminal vertexes. Consider the vertex energy function $\theta(c_j)$. If $\theta(c_j^0) < \theta(c_j^1)$, it indicates that candidate c_j is more likely to be 0. Thus we add an edge $(\hat{s}, c_j)$ from $\hat{s}$ to c_j and record a weight $w_j^{\hat{s}} = \theta(c_j^1) - \theta(c_j^0)$, as is shown in Fig. 3(b). Otherwise, we add edge $(c_j, \hat{t})$ from c_j to $\hat{t}$ and record a weight $w_j^{\hat{t}} = \theta(c_j^0) - \theta(c_j^1)$, as is shown in Fig. 3(c). Furthermore, we notice that the edge energy function also affects the connections between c_j and terminal vertexes. For each edge which goes out from c_j, e.g., (c_j, c_k), $j < k$, we compute $w = \theta(c_j^1, c_k^0) - \theta(c_j^0, c_k^0)$. If $w > 0$, it also indicates that candidate c_j is more likely to be 0. Then we connect $\hat{s}$ with c_j and set the weight as $w_j^{\hat{s}} = w$. If there is already an edge $(\hat{s}, c_j)$, then we update its weight as $w_j^{\hat{s}} = w_j^{\hat{s}} + w$. If $w < 0$, then we connect c_j with $\hat{t}$ and set the weight as $w_j^{\hat{t}} = -w$. Similarly, if there is already an edge $(c_j, \hat{t})$, then we update its weight as $w_j^{\hat{t}} = w_j^{\hat{t}} - w$. For each edge which goes to c_j, e.g., $(c_i, c_j), i < j$, we compute $w = \theta(c_i^1, c_j^1) - \theta(c_i^1, c_j^0)$. We check whether $w > 0$, and repeat the same process as mentioned above.

4.4 Min-cut Based Graph Separation

After we obtain the directed graph, as the last step of our method, we cut the graph to separate vertexes into two partitions. It is proved in literature [5] that the minimum cut problem is equivalent to the maximum flow problem from a source vertex to a sink vertex. In this work, we utilize the widely-used Edmonds-Karp (EK) algorithm [20] to cut the graph. Briefly, EK utilizes Breadth-First-Search (BFS) to find augmented paths iteratively and then updates the weights along the path until there is no new augmented path. When EK stops, we remove all edges whose weights are 0s and all vertexes that can be reached from $\hat{s}$. Last, the remaining vertexes, i.e., candidates that can be reached from $\hat{t}$, are our inferred truths.

4.5 MTD-CC Algorithm

To this end, we summarize the above mentioned procedures into Algorithm 1, named as MTD-CC, i.e., Multi-Truth Discovery with Candidate Correlations. The main body of the algorithm runs iteratively. For each iteration, we first store T to T', increase the counter, and reset the value of T to be empty. Then, we construct an MRF according to the method presented in Sect. 4.1 and 4.2. Also, we transform G into a directed graph $\hat{G}$ according to rules presented in Sect. 4.3. Next, we separate the graph $\hat{G}$ into two partitions based on the Min-cut theorem. Only those candidates who can be reached from the sink vertex $\hat{t}$ are regarded as truths. Last, we apply Eq. (1) to update all sources' reliabilities which will be used to compute truths in the next iteration. This algorithm converges when inferred truths of the current iteration agree with previous truths or the number of iterations exceeds a predefined threshold θ. It is worth noting that this algorithm only outputs truths obtained in the last iteration while all previous truths are used for updating sources' reliabilities.

Algorithm 1. MTD-CC

Input: A set of sources S, a set of candidates C.
Output: Inferred truths T, sources' reliabilities R.
1: Initialize r_i for all $s_i \in S$, counter$= 0$, $T = \emptyset$, $\theta \leftarrow$ predefined value.
2:**Repeat**
3: $T^{'} = T$, counter++.
4: $T = \emptyset$.
5: Construct an MRF $G(U, E)$.
6: Transform $G(U, E)$ into a directed graph $\hat{G}(\hat{U}, \hat{E})$.
7: Cut $\hat{G}$ using EK.
8: $T = T \cup \{c_i\}$ if c_i can be reached from $\hat{t}$.
9: Update all r_i using eq.(1).
10: **Until** $T == T^{'}$ or counter $>= \theta$.
11: **Return** T, R.

5 Performance Evaluation

5.1 Metrics and Baselines

Metrics. In order to evaluate the performance of our proposed algorithm, we take three general metrics which are often used for evaluating classification algorithms.

- *Precision.* The number of inferred truths which agree with the ground truths over the number of inferred truths.
- *Recall.* The number of inferred truths which agree with the ground truths over the number of ground truths.
- F_1 *score.* The harmonic mean of precision and recall, i.e., $F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$.

Baselines. We compare MTD-CC with five truth and multi-truth discovery methods, which are summarized as follows:

- *Majority Voting (MV).* For each candidate, it is regarded as a truth if more than a half of sources vote for it.
- *Truth Finder (TF)* [19]. TF investigates the relationship among different sources' answers, where each answer may include multiple candidates. It computes the score for each answer and the truth could be the answer with the highest score.
- *2-Estimates* [6]. It updates the weights of candidates and sources' reliabilities iteratively. One candidate is regarded as the truth if its weight exceeds a threshold.
- *LTM* [21]. LTM constructs a probabilistic graph to model the relations between truths and sources' reliabilities. It uses a sampling method to estimate the probability that a candidate is the truth. Those candidates whose probabilities exceed a threshold are treated as the truths.

– *DART* [12]. The Bayesian theorem is used in this method to update the scores of candidates and sources' reliabilities. A predefined threshold is also used to determine whether a candidate can be the truth.

To ensure the fair comparison, we run a series of experiments to warm up the initial parameter settings of baselines. As for our method, we initialize the source reliability as 0.6.

5.2 Experiment on Real Dataset

In order to evaluate MTD-CC and baseline methods in real applications, we run simulations on the book-author dataset [19]. This dataset includes 33,971 book-author records crawled from www.abebooks.com. Each record represents a bookstore's claim about the authors of a book. After removing redundant and invalid claims, we obtain 10,013 distinctive claims which are collected from 499 sources who report the authors of 483 books. We further select 100 books which we know the ground truths from the 483 books. We regard each book as an individual object and all performance results are presented by their averages.

Table 4. Comparison with different algorithms on the book-author dataset. The best performance values are in bold.

Methods	Book-author dataset		
	Precision	Recall	F_1 score
Majority Voting	0.883	0.633	0.738
Truth Finder	**0.9**	0.681	0.775
2-Estimates	0.867	0.675	0.759
LTM	0.803	0.856	0.828
DART	0.822	0.783	0.802
MTD-CC	0.825	**0.897**	**0.861**

Table 4 shows the performance of different algorithms on the book-author dataset in terms of precision, recall and F_1 score. We can see that MTD-CC achieves the best recall and F_1 score among all baseline methods. The reasons are two folds. First, the vertex and edge potential functions give advantages to our method to discover some truths which only have a few votes. For example, assume that the candidate c_j obtain 20 votes and c_k receives only 5 votes. If 4 votes for c_k also support c_j simultaneously, then c_k is more likely to be determined as a truth by MTD-CC. Second, given a certain object, the value of recall only depends on the number of inferred truths that agree with ground truths, so a method which discovers more ground truths will outperform other methods which discover less ground truths. This also explains why MTD-CC only achieves average result on the precision. Considering the overall performance, including F_1 score, the MTD-CC is still better than the baseline methods.

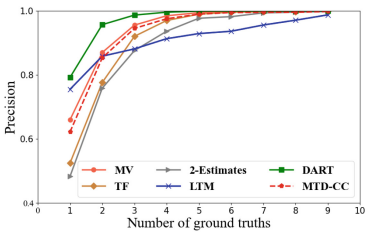
We can also see from Table 4 that the difference between precision and recall could be very large, e.g., Majority Voting, Truth Finder, and 2-Estimates. This shows that they omit some ground truths, even though their inferred truths are very likely to be correct. On the contrary, MTD-CC, along with LTM and DART, has a balance between the precision and recall. Among these three methods, MTD-CC is also the best on all metrics.

5.3 Experiment on Synthetic Dataset

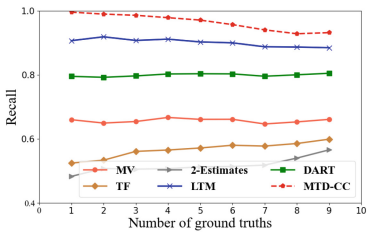
To further evaluate the performance of MTD-CC, we conduct experiments on a synthetic dataset. We investigate the influence of two factors: the number of ground truths and the number of candidates.

In the following experiments, we generate 100 sources. Each source is assigned with ground truths. However, in order to simulate the diversity of different sources, we allow each source to add noise to their votes. Specifically, when obtaining the ground truths, a source generates two probabilities, denoted as p_0 and p_1, where p_0 represents the probability of switching 0 to 1, and p_1 indicates the reverse switch.

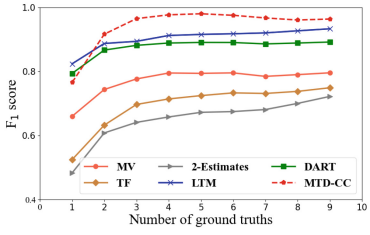
Effect of the Number of Ground Truths. We evaluate how the performance varies with different number of ground truths. We fix the number of candidates to be 10 in this experiment. The number of ground truths ranges from 1 to 9, and the corresponding truths are randomly selected from these 10 candidates.



(a) Precision.



(b) Recall.



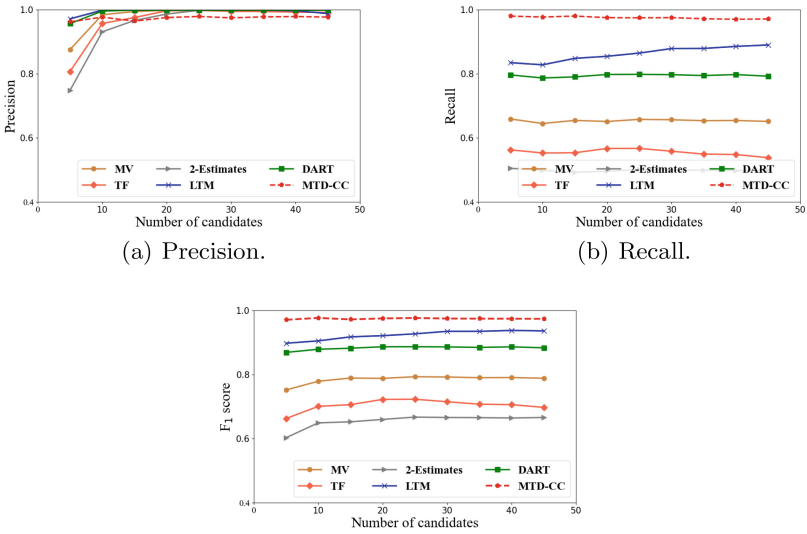
(c) F_1 score.

Fig. 4. Comparison of precision, recall and F_1 score on the synthetic dataset under different number of ground truths.

Investigating Fig. 4(a), we find that all methods show poor performance when the number of ground truths is 1, which corresponds to the single truth discovery. In this case, MTD-CC cannot guarantee to discover the ground truth precisely. In fact, MTD-CC often infers two truths. If one of them hits the ground truth, then the precision is 0.5.

Figure 4(b) shows that no matter how many ground truths, MTD-CC always shows the best recall. This result validates that the performance of MTD-CC is adaptable with the number of ground truths. Meanwhile, Fig. 4(c) shows that MTD-CC maintains the best F_1 score except the number of ground truth 1.

Effect of the Number of Candidates. We evaluate how the number of candidates to affect the performance of MTD-CC. We increase the number of candidates from 5 to 45, and set the number of ground truths to be half of the number of candidates. At this time, we can see from Fig. 5 that MTD-CC is always compelling and stable on all three metrics. We notice that there are 2 ground truths when the number of candidates is 5. At this point, the precision in Fig. 5(a) is much better than that in Fig. 4(a). The reason lies in the ratio of the number of ground truths to the number of candidates. We can see that the ratio is 2/5 in Fig. 5(a) and 2/10 in Fig. 4(a), which means all methods have more chance to infer ground truths.



(a) Precision. (b) Recall.

(c) F_1 score.

Fig. 5. Comparison of precision, recall and F_1 score on the synthetic dataset under different number of candidates.

6 Conclusion

In this paper, we study the problem of how to discover multiple truths from different sources' data. Our intuition is to improve the inference accuracy by utilizing correlations among different candidates. We design a metric to measure correlations between each pair of candidates based on sources' reliabilities and votes. Then, we use these correlations to construct an MRF and transform it into a directed graph with two additional terminal vertexes. Next, we separate the graph into two partitions based on the Min-cut theorem so that the truths can be directly inferred. Finally, experiment results on both real and synthetic datasets demonstrate that MTD-CC is notably more accurate than baseline methods.

Acknowledgements. This work was partially supported by the National Natural Science Foundation of China under Grants 62072061 and U20A20176, and by the Fundamental Research Funds for the Central Universities under Grant 2021CDJQY-026.

References

1. Dong, X., et al.: Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 601–610 (2014)
2. Du, Y., et al.: Bayesian co-clustering truth discovery for mobile crowd sensing systems. IEEE Trans. Industr. Inf. **16**(2), 1045–1057 (2019)
3. Du, Y., Xu, H., Sun, Y.-E., Huang, L.: A general fine-grained truth discovery approach for crowdsourced data aggregation. In: Candan, S., Chen, L., Pedersen, T.B., Chang, L., Hua, W. (eds.) DASFAA 2017. LNCS, vol. 10177, pp. 3–18. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-55753-3_1
4. Fang, X.S., Sheng, Q.Z., Wang, X., Chu, D., Ngu, A.H.: SmartVote: a full-fledged graph-based model for multi-valued truth discovery. World Wide Web **22**(4), 1855–1885 (2019)
5. Ford, L.R., Jr., Fulkerson, D.R.: Flows in Networks, vol. 54. Princeton University Press, Princeton (2015)
6. Galland, A., Abiteboul, S., Marian, A., Senellart, P.: Corroborating information from disagreeing views. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 131–140 (2010)
7. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press, Cambridge (2009)
8. Kolmogorov, V., Zabin, R.: What energy functions can be minimized via graph cuts? IEEE Trans. Pattern Anal. Mach. Intell. **26**(2), 147–159 (2004)
9. Li, Q., et al.: A confidence-aware approach for truth discovery on long-tail data. Proc. VLDB Endowment **8**(4), 425–436 (2014)
10. Li, Y., et al.: Reliable medical diagnosis from crowdsourcing: discover trustworthy answers from non-experts. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, pp. 253–261 (2017)
11. Li, Y., et al.: A survey on truth discovery. ACM SIGKDD Explor. Newsl. **17**(2), 1–16 (2016)
12. Lin, X., Chen, L.: Domain-aware multi-truth discovery from conflicting sources. Proc. VLDB Endowment **11**(5), 635–647 (2018)

13. Liu, T., Wu, W., Zhu, Y., Tong, W.: Accuracy-guaranteed event detection via collaborative mobile crowdsensing with unreliable users. In: Wang, X., Gao, H., Iqbal, M., Min, G. (eds.) CollaborateCom 2019. LNICST, vol. 292, pp. 729–744. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30146-0_49
14. Ma, F., et al.: Unsupervised discovery of drug side-effects from heterogeneous data sources. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 967–976 (2017)
15. Wang, X., Sheng, Q.Z., Fang, X.S., Yao, L., Xu, X., Li, X.: An integrated Bayesian approach for effective multi-truth discovery. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 493–502 (2015)
16. Wang, X., et al.: Truth discovery via exploiting implications from multi-source data. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 861–870 (2016)
17. Yang, Y., Bai, Q., Liu, Q.: A probabilistic model for truth discovery with object correlations. Knowl. Based Syst. **165**, 360–373 (2019)
18. Ye, C., et al.: Constrained truth discovery. IEEE Trans. Knowl. Data Eng. (2020)
19. Yin, X., Han, J., Philip, S.Y.: Truth discovery with multiple conflicting information providers on the web. IEEE Trans. Knowl. Data Eng. **20**(6), 796–808 (2008)
20. Zadeh, N.: Theoretical efficiency of the Edmonds-Karp algorithm for computing maximal flows. J. ACM (JACM) **19**(1), 184–192 (1972)
21. Zhao, B., Rubinstein, B.I., Gemmell, J., Han, J.: A Bayesian approach to discovering truth from conflicting sources for data integration. arXiv preprint arXiv:1203.0058 (2012)
22. Zheng, M., Cui, L., He, W., Guo, W., Lu, X.: A dynamic difficulty-sensitive worker distribution model for crowdsourcing quality management. In: Wang, X., Gao, H., Iqbal, M., Min, G. (eds.) CollaborateCom 2019. LNICST, vol. 292, pp. 12–27. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30146-0_2