World Wide Web (2019) 22:1855–1885
https://doi.org/10.1007/s11280-018-0629-3

SmartVote: a full-fledged graph-based model for multi-valued truth discovery

Xiu Susie Fang[1] · Quan Z. Sheng[1] · Xianzhi Wang[2] · Dianhui Chu[3] · Anne H. H. Ngu[4]

Received: 21 July 2017 / Revised: 20 May 2018 / Accepted: 30 July 2018 /
Published online: 22 August 2018

Abstract
In the era of Big Data, *truth discovery* has emerged as a fundamental research topic, which estimates data veracity by determining the reliability of multiple, often conflicting data sources. Although considerable research efforts have been conducted on this topic, most current approaches assume only one true value for each object. In reality, objects with multiple true values widely exist and the existing approaches that cope with multi-valued objects still lack accuracy. In this paper, we propose a full-fledged graph-based model, *SmartVote*, which models two types of source relations with additional quantification to precisely estimate source reliability for effective multi-valued truth discovery. Two graphs are constructed and further used to derive different aspects of source reliability (i.e., *positive precision* and *negative precision*) via random walk computations. Our model incorporates four important implications, including *two types of source relations*, *object popularity*, *loose mutual exclusion*, and *long-tail phenomenon on source coverage*, to pursue better accuracy in truth discovery. Empirical studies on two large real-world datasets demonstrate the effectiveness of our approach.

Keywords Truth discovery · Multi-valued objects · Graph-based model · Source relations · Object popularity · Long-tail phenomenon

1 Introduction

In today's digital and connected world, we are experiencing more and more freely created and published data on open sources. These massive data on the Web hold the potential to revolutionize many aspects of our modern society. For example, enterprises can leverage these data to analyze the market and promote their products. Government agencies can analyze

✉ Xiu Susie Fang
xiu.fang@students.mq.edu.au

Quan Z. Sheng
michael.sheng@mq.edu.au

Extended author information available on the last page of the article.

these data for decisions on security issues. Researchers can study these data for effective knowledge discovery. However, different sources often provide *conflicting* descriptions on the same objects of interest, due to typos, out-of-date data, missing records, or erroneous entries [10, 14, 16, 18, 28, 33], making it difficult to determine which data source to trust. Those conflicting data could mislead people and cause considerable damages and financial loss in various applications such as drug recommendation in healthcare systems or price prediction in the stock markets [1]. Moreover, due to the massiveness of data, it is infeasible to manually determine which data is true.

Considerable research efforts have been conducted to address the truth discovery problem [5, 17, 23, 30, 41, 46]. These methods commonly assume that each object has exactly one true value (i.e., *single-valued* assumption). However, multi-valued objects widely exist in the real world, e.g., the children of a person, the authors of a book. One may argue that the previous *single-valued methods*) (i.e., the methods under the single-value assumption) can be adapted to deal with multi-valued objects—given a multi-valued object, these methods can simply regard the value set claimed by each source as a joint single value, and then determine the most confident value set as the truth. However, the value sets provided by different sources might be correlated. For example, there may be some overlaps between two sources' claimed value sets, indicating that they are not totally voting against each other. Neglecting this implication could degrade the accuracy of truth discovery. Moreover, single-valued methods overlook the important distinction between two aspects of quality: false negatives and false positives. Another example is that, for multi-valued objects, some sources may provide erroneous values, making false positives, while some other sources may provide partial true values without erroneous values, making false negatives. Single-valued truth discovery methods cannot distinguish the quality of these two types of sources.

In view of the limitations of single-valued methods, several approaches [35–38, 46] have been proposed to tackle the multi-valued objects. However, these approaches have disadvantages and the multi-truth discovery (MTD) problem is still far from being solved. Firstly, previous methods generally require proper initialization of source reliability, which significantly impacts their performance in terms of convergence rate and accuracy. Precise source reliability initialization with limited parameter settings could facilitate better truth discovery. Secondly, there are *supportive relations* among sources, implying that sources agree with/endorse another source by providing the same values. Intuitively, a source endorsed by more sources is regarded as more authoritative and can be more trusted. Unlike other source relations (e.g., copying relations) that have been widely studied, the supportive relations among sources are largely neglected by the previous work. Thirdly, while *object difficulty* [17] (i.e., the difficulty of getting true values for varied objects) and *relations among objects* [30, 43] (i.e., objects may have relations that affect each other) have been studied by previous research efforts, to the best of our knowledge, no previous work differentiate the popularity of different objects. In reality, the impact of knowing the true values of various objects might be totally different. For example, between the email address and the children of a famous researcher, her email address is apparently more popular because other researchers doing research in the same areas may need the email address to contact her. Therefore, taking object popularity into consideration could better model the real-world truth discovery and therefore lead to the more accurate result.

In a nutshell, this work makes the following main contributions:

- We measure two aspects of source reliability, i.e., *positive precision* and *negative precision* to differentiate the false positives and false negatives. We propose a graph-based approach, called *SourceVote*, to improve the existing truth discovery methods. Two graphs are constructed by capturing two-sided inter-source agreements, from which the source reliability is derived.

- We propose a graph-based model, called *SmartVote*, as an overall solution to the multi-truth discovery problem. This model incorporates four important implications, including *two types of source relations*, *object popularity*, *loose mutual exclusion*, and *long-tail phenomenon on source coverage*, for better truth discovery.
- By relaxing the assumption that sources are independent of each other, we model two types of source relations, namely *supportive relations* and *copying relations*. Graphs capturing source features are constructed based on these relations. Specifically, source authority features and two-sided source precision are captured by ±*supportive agreement graphs*, while source dependence scores are quantified by ±*malicious agreement graphs*. Random walk computations are applied to both types of graphs to estimate source reliability and dependence scores.
- We propose to differentiate the popularity of objects by leveraging object occurrences and source coverage, to minimize the misleading of false values. The long-tail phenomenon on source coverage is not rare in the real world. Our model tackles MTD by considering this phenomenon and avoids the quality of sources with very few claims from being under (or over) estimated.
- We conduct extensive experiments to demonstrate the effectiveness of our approach via comparison with competitive state-of-the-art baseline methods on two real-world datasets. The impact of different implications on our model is also empirically studied and discussed.

The rest of the paper is organized as follows. We discuss the observations that motivate our work and formulate the multi-valued truth discovery problem in Section 2. Section 3 introduces the SourceVote approach for improving the existing truth discovery methods. Section 4 presents the model of SmartVote and the incorporated implications. We report our experimental results in Section 5, and discuss the related work in Section 6. Finally, Section 7 provides some concluding remarks.

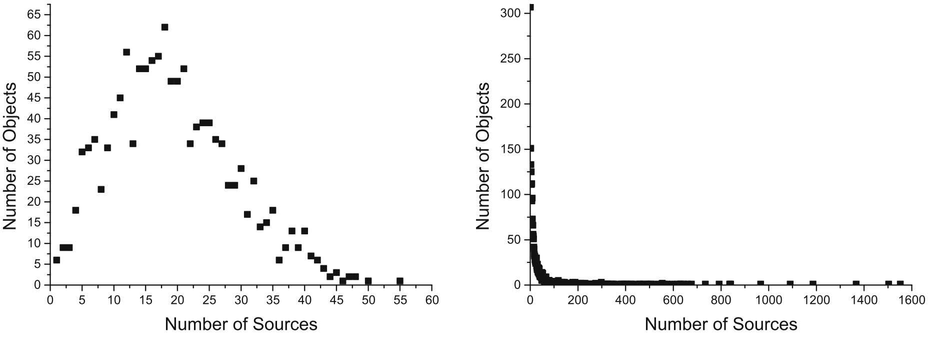
2 Preliminaries

We present the statistical observations of objects on real-world datasets in Section 2.1 to motivate our approach. In Section 2.2, we formally define the multi-valued truth discovery problem. We further validate the intuition that motivates us to model source reliability by quantifying the two-sided inter-source agreements in Section 2.3.

2.1 Observations and motivations for object differentiation

We have investigated the distributions of objects over sources in various real-world datasets. As an example, Figure 1a and b show the results on the *Book-Author* [41] and *Biography*[1] [30] datasets, respectively. Each point (x, y) in the figure depicts that y objects are covered by x sources, respectively, in the corresponding dataset. We observe an apparent long-tail phenomenon from the distributions of the Biography dataset (containing 2,579 objects), which indicates that very few objects are referenced by a large number of sources

[1] In this paper we focus on the parent-children relation in the dataset because it corresponds to multi-valued objects.



(a) Book-Author Dataset

(b) Biography Dataset

Figure 1 The number of sources that provide values on objects: different objects are covered by varying number of sources

in the dataset, and many objects are covered by very few sources. For the Book-Author dataset with much fewer objects (containing 1,262 objects), the long-tail phenomenon is less evident, but objects are claimed by significantly varying numbers of sources, indicating that objects are of different occurrences. For example, there are 624 sources in total in the Book-Author dataset. The author list of book (id : 1558606041) is claimed by 55 sources, while the lists of books (id : 0201608359) and (id : 020189551X) are only claimed by one source each.

Intuitively, the objects with more occurrences in the sources' claims indicate that they are more popular and sources tend to publish popular information to gain attention from the public. Since the number of potential audiences of popular objects is usually bigger than that of unpopular objects, if a source provides false values on a popular object, it will mislead more people than on a less popular object. With this consideration, we believe that there is a different impact on the public for knowing the true values of different objects. Therefore, we propose to distinguish source reliability by differentiating the *popularity* of objects, to minimize the number of people misguided by false values. Sources providing false values for popular objects should be penalized more and assigned with lower reliability, to discourage them from misguiding the public. Therefore, more evidence can be used for estimating value veracity regarding those objects, leading to reliable truth estimation. This motivates us to assign heavier weights to popular objects in the calculation of source reliability for more accurate estimation.

2.2 Problem definition

A multi-truth discovery problem (i.e., MTD) generally involves five components (Table 1 summarizes the notations used in this paper) in its life cycle:

Explicit inputs include i) a set of multi-valued *objects*, $\mathcal{O}$, each of which may have more than one true value to be discovered. The number of true value(s) can vary from object to object; ii) a set of *sources*, $\mathcal{S}$. Each $s \in \mathcal{S}$ provides potential true values on a subset of objects in $\mathcal{O}$; iii) *claimed values*, the values provided by any source of $\mathcal{S}$ on any objects of $\mathcal{O}$. Given a source s, we regard the set of values provided by s on an object o as *positive claims*, denoted as $\mathcal{V}_{s_o}$.

Table 1 Notations used in the paper

Notation	Explanation
$o, \mathcal{O}$	An object (resp., set of all objects)
$s, \mathcal{S}$	A source (resp., set of all sources)
$v, \mathcal{V}$	A claimed value (resp., set of all claimed values)
$\mathcal{V}_o{}^*$	Identified truth for o
$\mathcal{V}_o{}^g$	Ground truth for o
$\mathcal{S}_o$	Set of sources provide values on o
$\mathcal{S}_v, \mathcal{S}_{\tilde{v}}$	Set of sources claim/disclaim v on o
$\mathcal{O}_s$	Set of objects covered by s
$Cov(s)$	The coverage of s
$\mathcal{V}_{s_o}, \tilde{\mathcal{V}}_{s_o}$	Set of positive/negative claims provided by s on o
$\mathcal{U}_o$	Set of all claimed values on o
$\tau(s), \tilde{\tau}(s)$	±SourceVote, evaluation of positive/negative precision of s
$\tau'(s), \tilde{\tau}'(s)$	±SmartVote, improved evaluation of positive/negative precision of s
$\mathcal{C}_v, \mathcal{C}_{\tilde{v}}$	The confidence score of v being true/false
$\mathcal{A}(s_1, s_2), \tilde{\mathcal{A}}(s_1, s_2)$	Endorsement degree from s_1 to s_2 on positive/negative claims
$A_o(s_1, s_2), \tilde{A}_o(s_1, s_2)$	Agreement between the positive/negative claims of s_1 and s_2 on o
$\mathcal{P}_o$	The popularity degree of o
$\mathcal{D}(s, o), \tilde{\mathcal{D}}(s, o)$	The dependence score of s providing positive/negative claims on o
$\mu(s, o), \tilde{\mu}(s, o)$	The confidence score of s providing positive/negative claims on o
$\mathcal{L}(s_1, s_2)$	The long-tail phenomenon compensation for edge from s_1 to s_2
$\omega(s_1 \rightarrow s_2), \tilde{\omega}(s_1 \rightarrow s_2)$	The weight of edge from s_1 to s_2 in ±agreement graph
$\omega'(s_1 \rightarrow s_2), \tilde{\omega}'(s_1 \rightarrow s_2)$	The weight of edge from s_1 to s_2 in ±supportive agreement graph
$\omega_{c_o}(s_1 \rightarrow s_2), \tilde{\omega}_{c_o}(s_1 \rightarrow s_2)$	The weight of edge from s_1 to s_2 in ±malicious agreement graph of o

Implicit inputs are derived from the explicit inputs and include two parts: i) the complete set of values provided by all sources on any object o, denoted as $\mathcal{U}_o$; ii) by incorporating the *mutual exclusion assumption*, given an object o, a source s that makes positive claims $\mathcal{V}_{s_o}$ is believed to implicitly disclaim all the other values on o. We denote the set of values disclaimed by s as $\tilde{\mathcal{V}}_{s_o}$ (i.e., *negative claims* provided by s on o), which is calculated by $\mathcal{U}_o - \mathcal{V}_{s_o}$.

Intermediate variables are generated and updated during the iterative truth discovery procedure. They include i) *source reliability*, which reflects the capability of each source providing true values; ii) *confidence score*, which reflects the confidence on a value being true or false. In this paper, we differentiate the false positives and false negatives made by sources by modeling two aspects of source reliability, namely positive precision (i.e., the probability of the positive claims of a source being true), and negative precision (i.e., the probability of the negative claims of a source being false). In the following sections, we will first derive ±*SourceVote* (denoted by $\tau(s)$ and $\tilde{\tau}(s)$ for a source s) as the evaluations of source reliability by capturing source authority features based on two-sided inter-source agreements (Section 3). Then, we improve the source reliability evaluations by incorporating four important implications in Section 4. The improved evaluations of source positive and negative precision are named as ±*SmartVote*, denoted by $\tau'(s)$ and $\tilde{\tau}'(s)$ for a source s.

Table 2 An illustrative example: four sources provide author names of two books

	9780072830613	9780072231236
Ground Truth	Stephen;James	Michael
s_1	Stephen;James;Merrill	Michael;Lloyd
s_2	Stephen;James	Michael
s_3	Stephen;Kate	Michael;Susan
s_4	Stephen;Kate	Michael;Susan

Accordingly, we estimate both the confidence scores of a value v being true (i.e., $\mathcal{C}_v$) and false ($\mathcal{C}_{\tilde{v}}$).

Outputs are the *identified truth* for each object $o \in \mathcal{O}$, denoted as $\mathcal{V}_o{}^*$.

Ground truth is the factual truth for each object $o \in \mathcal{O}$, denoted as $\mathcal{V}_o{}^g$, which is used to measure the effectiveness of the truth discovery methods.

Example 1 Table 2 shows a snippet of the Book-Author dataset. In this particular example, four sources (i.e., s_1, s_2, s_3, and s_4) claim values on two objects (i.e., the authors of two books id : 9780072830613 and id : 9780072231236, denoted as o_1 and o_2). Each cell in the table demonstrates the positive claims of a specific source on a specific object. For example, s_1 provides {Stephen;James;Merrill} as positive claims, i.e., $\mathcal{V}_{s_{o_1}}$, on object o_1. There are conflicts among these four sources as they provide different positive claims on the same objects. Table 2 also shows the ground truth of the two objects. Given the conflicting data, our goal is to identify the true authors for these two books. We can derive from the dataset that $\mathcal{U}_{o_1}$ = {Stephen;James;Merrill;Kate}, $|\mathcal{U}_{o_1}| = 4$, and $\mathcal{U}_{o_2}$ = {Michael;Lloyd;Susan}, $|\mathcal{U}_{o_2}| = 3$, based on which we can further extract implicit inputs regarding the two objects from the raw dataset as shown in Tables 3 and 4. Comparing with the ground truth, s_2 provides all the true values, which deserves a higher positive precision and negative precision. Sources s_3 and s_4 provide the same values and there may be supportive relations or copying relations between them. We will discuss these two types of relations in Section 4. Source s_1 is audacious, which claims additionally a false value besides the true values for every object, while s_3 and s_4 are error-prone, each of which claims a false value for every object.

We formally define the multi-valued truth discovery problem as follows:

Table 3 Truth discovery inputs regarding the first book

9780072830613				
	positive claims (pc)	# pc	negative claims (nc)	# nc
s_1	Stephen;James;Merrill	3	Kate	1
s_2	Stephen;James	2	Merrill;Kate	2
s_3	Stephen;Kate	2	James;Merrill	2
s_4	Stephen;Kate	2	James;Merrill	2

Table 4 Truth discovery inputs regarding the second book

9780072231236				
	positive claims (pc)	# pc	negative claims (nc)	# nc
s_1	Michael;Lloyd	2	Susan	1
s_2	Michael	1	Lloyd;Susan	2
s_3	Michael;Susan	2	Lloyd	1
s_4	Michael;Susan	2	Lloyd	1

Definition 1 Multi-Truth Discovery Problem (MTD) Given a set of multi-valued objects ($\mathcal{O}$) and a set of sources ($\mathcal{S}$) that provide conflicting values $\mathcal{V}$, the goal of MTD is to identify a set of true values (${\mathcal{V}_o}^*$) from $\mathcal{V}$ for each object o, satisfying that ${\mathcal{V}_o}^*$ is as close to the ground truth ${\mathcal{V}_o}^g$ as possible. A truth discovery process often proceeds along with the estimation of the reliability of sources, i.e., positive precision ($\tau'(s)$) and negative precision ($\tilde{\tau}'(s)$). The perfect truth discovery results will satisfy ${\mathcal{V}_o}^* = {\mathcal{V}_o}^g$.

2.3 Agreement as hint

For multi-valued objects, sources may provide totally different, the same, or overlapping sets of values from one another. Given an object, we define the common values claimed by two sources on the object as *inter-source agreement*. Based on the mutual exclusion, we consider two-sided inter-source agreements. Specifically, *+agreement* (resp., *–agreement*) is the agreement between two sources on positive (resp., negative) claims, indicating that they agree with each other on their claimed (resp., disclaimed) common values being true (resp., false). Intuitively, the agreement among sources indicates an endorsement. If the positive (resp., negative) claims of a source are agreed/endorsed by the majority of other sources, this source may have a high positive (resp., negative) precision and is called an *authoritative source*.

Suppose ${\mathcal{V}_o}^g$ is the ground truth of an object o, $\mathcal{U}_o$ is the set of all claimed values of o, we denote by $\mathcal{U}_o - {\mathcal{V}_o}^g$ the set of false values of o. For the simplicity of presentation, we use T, U, and F to represent ${\mathcal{V}_o}^g, \mathcal{U}_o$, and $\mathcal{U}_o - {\mathcal{V}_o}^g$ in this section. For any two sources s_1 and s_2, the +agreement between them on an object o is calculated as:

$$A_o(s_1, s_2) = \mathcal{V}_{s_{1_o}} \cap \mathcal{V}_{s_{2_o}} \tag{1}$$

Suppose s_1 and s_2, each selects a true value from T independently. We denote their selected values as t_1 and t_2, respectively. The probability of $t_1 = t_2$, denoted as $P_{A_o}(t_1, t_2)$, can be calculated as follows[2]:

$$P_{A_o}(t_1, t_2) = \frac{1}{|T|} \tag{2}$$

Similarly, let f_1 and f_2 be the two values independently selected by s_1 and s_2 from F, and $P_{A_o}(f_1, f_2)$ be the probability of s_1 and s_2 providing the same false value (i.e., $f_1 = f_2$), $P_{A_o}(f_1, f_2)$ can be calculated using:

$$P_{A_o}(f_1, f_2) = \frac{1}{|F|} \tag{3}$$

[2] Note that this probability is based on a prior knowledge that s_1 and s_2 each provides a true value, which is different from the probability of two sources s_1 and s_2 independently provide the same true value.

In reality, an object usually has a small truth set and random false values, i.e., $|T| \ll |U|$. Applying this to (2) and (3), we get:

$$P_{A_o}(f_1, f_2) \ll P_{A_o}(t_1, t_2) \tag{4}$$

Typically, the values claimed by sources would contain a fraction of true values from T and a faction of false values from F. By applying (4), positive claims from T are more likely to agree with each other than those from F. This implies that the more true values a source claims, the more likely the other sources agree with its claimed values. Inversely, if a source shows a high degree of agreement with the other sources regarding its claimed values, the values claimed by this source would have a higher probability to be true, and this source would have a bigger positive precision. The positive precision of a source is endorsed by the +agreements between this source and the other sources.

Similarly, the –agreement between any two sources s_1 and s_2 on an object o is calculated as:

$$\tilde{A}_o(s_1, s_2) = \tilde{\mathcal{V}}_{s_{1_o}} \cap \tilde{\mathcal{V}}_{s_{2_o}} = U - (\mathcal{V}_{s_{1_o}} \cup \mathcal{V}_{s_{2_o}}) \tag{5}$$

Let $\tilde{A}_o(s_1, s_2) \cap T$ be the true values in the –agreement between s_1 and s_2, and $\tilde{A}_o(s_1, s_2) \cap F$ be the false values in the –agreement between s_1 and s_2, satisfying $|\mathcal{V}_{s_{1_o}}| \ll |U|$, $|\mathcal{V}_{s_{2_o}}| \ll |U|$, $|T| \ll |U|$. It can be proved that $|\tilde{A}_o(s_1, s_2) \cap T| \ll |\tilde{A}_o(s_1, s_2) \cap F|$. Therefore, it is more likely for sources to agree with one another on false values than on true values with respect to their negative claims. This implies that the more false values a source disclaims, the more likely that other sources agree with its negative claims. Inversely, if a source shows a high degree of agreement with the other sources on its negative claims, the values disclaimed by this sources would have higher probabilities to be false, and this source would have a bigger negative precision. The negative precision of a source is endorsed by the –agreements between this source and the other sources.

3 The SourceVote approach

In this section, we propose a two-step process called *SourceVote* to evaluate the two-sided source reliability, which can be utilized to initialize the existing truth discovery methods: i) creating two graphs based on ±agreement among the sources, namely ±*agreement graph*s, and ii) assessing source reliability as ±SourceVote based on the graphs. We first assume sources as independent of one another and then relax this assumption in Section 4.

To facilitate source reliability evaluation, we model the two-sided inter-source agreements by constructing two fully connected weighted graphs, namely ±*agreement graph*s. In such a graph, the vertices denote sources, each directed edge represents that one source agrees with the other source, and the weight on each edge depicts to what extent one source endorses the other source. We define $\mathcal{A}(s_1, s_2)$ (resp., $\tilde{\mathcal{A}}(s_1, s_2)$) as the *endorsement degree* from s_1 to s_2 on positive (resp., negative) claims, representing the rate, at which s_2 is endorsed by s_1 on the values being true (resp., false).

+Agreement graph We first formalize endorsement between any two sources based on their common positive claims as follows:

$$\mathcal{A}(s_1, s_2) = \sum_{o \in \mathcal{O}_{s_1} \cap \mathcal{O}_{s_2}} \frac{|A_o(s_1, s_2)|}{|\mathcal{V}_{s_{2_o}}|} \tag{6}$$

We calculate the weight on the edge from s_1 to s_2 using:

$$\omega(s_1 \to s_2) = \beta + (1-\beta) \cdot \frac{\mathcal{A}(s_1, s_2)}{|\mathcal{O}_{s_1} \cap \mathcal{O}_{s_2}|} \tag{7}$$

In (7), we add a "*smoothing link*" by assigning a small weight to every pair of vertices, where β is the smoothing factor. This measure guarantees that the graph is always connected and source reliability calculation can converge. For our experiments, we set $\beta = 0.1$ (as demonstrated by Gleich et al. [19] in their empirical studies). Finally, we normalize the weights of out-going links from every vertex by dividing the edge weights by the sum of the out-going edge weights from the vertex. This normalization allows us to interpret the edge weights as the transition probabilities for the random walk computation.

–Agreement graph We construct the –agreement graph in a similar way by applying the following two equations:

$$\tilde{\mathcal{A}}(s_1, s_2) = \sum_{o \in \mathcal{O}_{s_1} \cap \mathcal{O}_{s_2}} \frac{|\tilde{A}_o(s_1, s_2)|}{|\tilde{\mathcal{V}}_{s_{2o}}|} \tag{8}$$

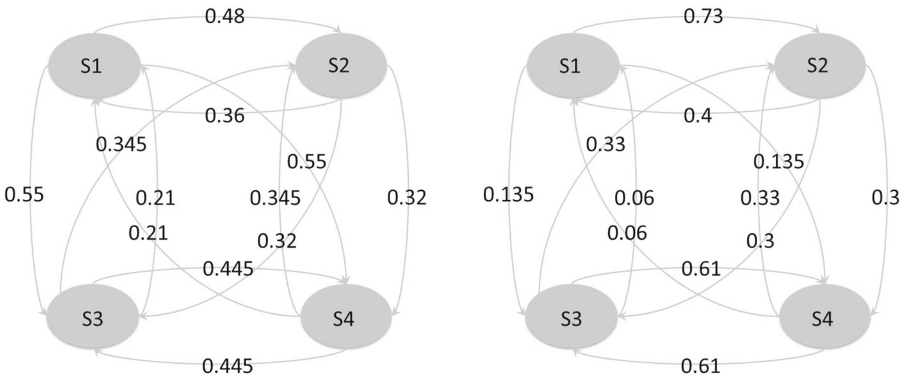
$$\tilde{\omega}(s_1 \to s_2) = \beta + (1-\beta) \cdot \frac{\tilde{\mathcal{A}}(s_1, s_2)}{|\mathcal{O}_{s_1} \cap \mathcal{O}_{s_2}|} \tag{9}$$

Specifically, we adopt the *Fixed Point Computation Model* (FPC) to calculate ±SourceVote for each source. FPC captures the transitive propagation of source trustworthiness through agreement links based on the above-constructed two graphs [4]. In particular, we refer to each graph as a Markov chain, with vertices as the states, and the weights on the edges as the probabilities of transition between the states. Then, we calculate the asymptotic stationary visit probabilities of the Markov random walk. As the sum of all visit probabilities equals to 1, they cannot reflect the real source precision. To resolve this issue, we set the positive precision (resp., negative precision) of the source with the highest visit probability in the +agreement graph (resp., –agreement graph) as pp_{max} (resp., np_{max}), and calculate the *normalization rate* by dividing the precision by the corresponding visit probability. Then, the visit probabilities of all sources can be normalized as precision (±SourceVote) (denoted as $\tau(s)$ and $\tilde{\tau}(s)$) by multiplying the normalization rate. The computed ±SourceVote capture the following two characteristics:

- Vertices with more input edges have higher precision since those sources are endorsed by a large number of sources and should be more trustworthy[3].
- Endorsement from a source with more input edges should be more trusted than that from other sources. Since an authoritative source is likely to be more trustworthy, the source endorsed by an authoritative source is also likely to be more trustworthy.

Example 2 Figure 2 shows the sample ±agreement graphs for the dataset described in Example 1. Take the link from s_1 to s_2 in the sample +agreement graph as an example, by applying (1), we get $|A_{o_1}(s_1, s_2)| = 2$, and $|A_{o_2}(s_1, s_2)| = 1$. By substituting this result into (6), $\mathcal{A}(s_1, s_2) = \frac{2}{2} + \frac{1}{1} = 2$, and by further substituting this result into Eq. 7, $\omega(s_1 \to s_2) = 0.1 + (1-0.1) \times \frac{2}{2} = 1$. In the same way, we obtain $\omega(s_1 \to s_3) = 0.55$, and $\omega(s_1 \to s_4) = 0.55$. Finally, the normalized weights of edges $s_1 \to s_2$, $s_1 \to s_3$, and

[3] Here we neglect the smoothing links, if there is no common value between two sources, there is no link between them in the graphs.



(a) +Agreement Graph (b) –Agreement Graph

Figure 2 Sample ±agreement graphs of four sources in Table 2

$s_1 \rightarrow s_4$ of +agreement graph are $\frac{1}{1+0.55+0.55} = 0.48$, $\frac{0.55}{1+0.55+0.55} = 0.26$, $\frac{0.55}{1+0.55+0.55} = 0.26$, respectively. After random walk computations, we obtain the visit probabilities of the sources in the +agreement graph as $\{s_1 : 0.21, s_2 : 0.28, s_3 : 0.26, s_4 : 0.26\}$, and those in the –agreement graph as $\{s_1 : 0.17, s_2 : 0.29, s_3 : 0.27, s_4 : 0.27\}$. Suppose the real positive precision of s_2 is 1, and the real negative precision of s_2 is also 1, we finally obtain the +SourceVote of the sources as $\{s_1 : 0.75, s_2 : 1, s_3 : 0.93, s_4 : 0.93\}$, and –SourceVote of them as $\{s_1 : 0.59, s_2 : 1, s_3 : 0.93, s_4 : 0.93\}$. We can see that the above results capture the authority features of those four sources. For example, the positive claims and negative claims of s_2 are both endorsed by more sources than those of the other sources. Thus, s_2 is assigned with the highest ±SourceVote. According to the ground truth provided by Table 2, s_3 and s_4 agree with each other on false values. This indicates a possible malicious agreement between them and, consequently, over-estimated positive precision and negative precision of s_3 and s_4 (this aspect will be described in detail in Section 4.2).

Note that the majority of existing truth discovery methods start with initializing source reliability as a default value (e.g., always set as 0.8 for source reliability [34]). These initialization requirements impact the convergence rate and precision of the methods. Li et al. state in [24] that "*knowing the precise trustworthiness of sources can fix nearly half of the mistakes in the best fusion results*". As the construction and computation of our agreement graphs can be easily realized, SourceVote can be applied to all existing methods for precise source reliability initialization. Specifically, +$SourceVote$ (resp., –$SourceVote$) is for positive (resp., negative) precision initialization.

4 The SmartVote approach

In reality, sources might not only support one another by providing the same true claims but also maliciously copy from others to provide the same false claims. Therefore, we further identify two types of source relations to conduct more accurate source reliability estimation. Specifically, sharing the same true values means one source supports/endorses the other source, indicating a *supportive relation* between the two sources. We define the common values between these two sources as *supportive agreement*. Based on the analysis in Section 2.3, we measure source reliability by quantifying inter-source supportive agreements. Even though one source can copy from another, we consider this type of copying relations as benignant. On the contrary, sharing the same false values is typically a rare event when the sources are fully independent. If two sources share a significant amount of false values, they are likely to copy from each other, indicating a *copying relation* between them. We define these common false values as *malicious agreement*. Neglecting the existence of deliberate copying of false values would impair the accuracy of source reliability estimation.

Besides source relations, several additional heuristics can also be considered to precisely estimate source reliability in reality. To this end, we propose a full-fledged graph-based model, called $SmartVote$, to solve the MTD problem.

4.1 The graph-based model

Given the large-scale noisy multi-sourced Web data, it is difficult for a human to determine what the truth is. The goal of our model is to automatically predict the truth from the conflicting multi-sourced data. Our model incorporates four implications, including source relations, object popularity, loose mutual exclusion, and long-tail phenomenon on source coverage, in a graph-based model. We classify and briefly describe the above components as follows:

Core component It applies two principles for truth discovery [26]: 1) sources providing more true values are assigned with higher reliability; 2) values provided by higher-quality sources are more likely to be true. Value confidence scores and source reliability are iteratively calculated from each other until convergence. By relaxing the source independent assumption and identifying two types of source relations (namely *supportive relations* and *copying relations*), the general inter-source agreements quantified by SourceVote are divided into *supportive agreements* and *malicious agreements*. The SmartVote core component derives the improved evaluations of source positive and negative precision, i.e., ±*SmartVote*, from two constructed ±*supportive agreement graphs*. The constructions of ±supportive agreement graphs incorporate the outputs of the four optimization components. Note that the supportive relations among sources are modeled by supportive agreement graphs constructed by the core component.

Optimization components These four optimization components compute the parameters regarding the four implications required by the core component. The *malicious agreement detection* component models the copying relations among sources and derives the dependence score of each source providing claims on each object (Section 4.2). The *object popularity quantification* component differentiates the popularity of objects based on the consideration that knowing the truths of different objects impacts differently on source reliability estimation (Section 4.3). Given a multi-valued object, a source may either cautiously provide partial true values while omitting the values they are not sure about, or audaciously provide all potential values even if the veracity of the claimed values is uncertain. This means that the mutual exclusion among values is not as strict as that of the single-valued object, i.e., the loose mutual exclusion. For this reason, SmartVote uses the *source confidence measurement* component to calculate the source confidence scores of providing positive (resp., negative) claims on each object, and reconcile sources' belief in their positive and negative claims (Section 4.4). Finally, the *balancing long-tail phenomenon on source* *coverage* component calculates the compensation of long-tail phenomenon on source coverage for each link in the ±supportive agreement graphs to avoid small sources from being assigned with extreme reliability (Section 4.5).

In the core component, the constructions of ±supportive agreement graphs are similar to those of ±agreement graphs. In particular, we calculate the endorsement degree from s_1 to s_2 on positive claims by modifying (6) as follows:

$$\mathcal{A}(s_1, s_2) = \mathcal{L}(s_1, s_2) + \sum_{o \in \mathcal{O}_{s_1} \cap \mathcal{O}_{s_2}} \frac{|A_o(s_1, s_2)|}{|\mathcal{V}_{s_{2o}}|} \cdot (1 - \prod_{v \in A_o(s_1, s_2)} \mathcal{C}_{\tilde{v}}) \cdot \mathcal{P}_o \cdot (1 - \mathcal{D}(s_1, o)) \cdot \mu(s_1, o) \tag{10}$$

where $\mathcal{D}(s, o)$ is the dependence score of s providing positive claims on o (defined in Section 4.2), $\mathcal{P}_o$ is the popularity degree of o (defined in Section 4.3), $\mu(s, o)$ is the confidence score of s providing positive claims on o (defined in Section 4.4), and $\mathcal{L}(s_1, s_2)$ is the long-tail phenomenon compensation of edge from s_1 to s_2 (defined in Section 4.5).

We calculate the edge weight of a +supportive agreement graph as follows:

$$\omega'(s_1 \rightarrow s_2) = \beta + (1 - \beta) \cdot \frac{\mathcal{A}(s_1, s_2)}{|\mathcal{O}_{s_2}|} \tag{11}$$

Similarly, we calculate the edge weight of a –supportive agreement graph as follows:

$$\tilde{\mathcal{A}}(s_1, s_2) = \mathcal{L}(s_1, s_2) + \sum_{o \in \mathcal{O}_{s_1} \cap \mathcal{O}_{s_2}} \frac{|\tilde{A}_o(s_1, s_2)|}{|\tilde{\mathcal{V}}_{s_{2o}}|} \cdot (1 - \prod_{v \in \tilde{A}_o(s_1, s_2)} \mathcal{C}_{v}) \cdot \mathcal{P}_o \cdot (1 - \tilde{\mathcal{D}}(s_1, o)) \cdot \tilde{\mu}(s_1, o) \tag{12}$$

$$\tilde{\omega}'(s_1 \rightarrow s_2) = \beta + (1 - \beta) \cdot \frac{\tilde{\mathcal{A}}(s_1, s_2)}{|\mathcal{O}_{s_2}|} \tag{13}$$

We apply FPC random walk to those two graphs and obtain $\tau'(s)$ and $\tilde{\tau}'(s)$ as the ±SmartVote for each source by conducting the same normalization process as with SourceVote. Besides the features captured by ±SourceVote, ±SmartVote additionally capture the following characteristics:

- The endorsement from a source on values with higher probabilities to be true (resp., false) in the +supportive agreement graph, should be more (resp., less) respected. Similarly, the endorsement from a source on values with higher probabilities to be false (resp., true) in the –supportive agreement graph, should be more (resp., less) respected.
- The endorsement independently provided by a source should be more trustworthy, since the endorsement provided by copiers can be malicious and might have little wisdom in it. Also, the endorsement from a source on popular objects should be highlighted, since popular objects are more valued by the public, false values of which can lead to bad consequences.
- If a source shows more confidence in the claims (positive or negative) of an object, the endorsement from this source on the object should be highlighted. Also, sources covering few objects should not be assigned with extremely big or small values of ±SmartVote as the evidence for estimating their reliability are limited.

To jointly determine value veracity from source reliability, we consider each source that belongs to $\mathcal{S}_o$ casts a smart vote to each potential value of o. In particular, if a source provides v as a positive claim, then it casts a vote proportional to $\tau'(s)$ for it; in contrast, if a source disclaims v, then it casts a vote proportional to $(1 - \tau'(s))$ for it. Finally, we compute the confidence scores of each value v being true or false by applying the following equations:

$$\mathcal{C}_v = \frac{\sum_{s \in \mathcal{S}_v} \tau'(s) + \sum_{s \in \mathcal{S}_{\tilde{v}}} (1 - \tilde{\tau}'(s))}{|\mathcal{S}_o|} \tag{14}$$

$$\mathcal{C}_{\tilde{v}} = \frac{\sum_{s \in \mathcal{S}_v} (1 - \tau'(s)) + \sum_{s \in \mathcal{S}_{\tilde{v}}} \tilde{\tau}'(s)}{|\mathcal{S}_o|} \tag{15}$$

4.2 Malicious agreement detection

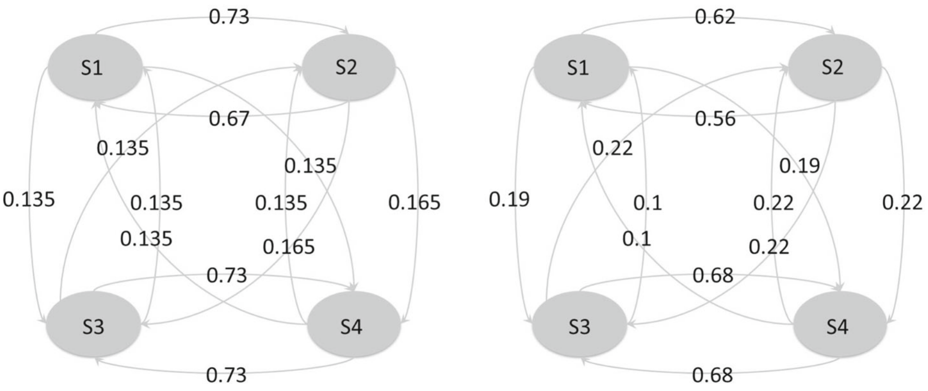
Copying relations among sources in the real world are complex. For example, a copier may copy all values or partial values from a source. A source may transitively copy from another source or collect information from several sources. Multiple sources may copy one source. To better model the malicious agreement among sources globally, we construct ±malicious agreement graphs for sources that provide values on an object o, i.e., $\mathcal{S}_o$, for each object $o \in \mathcal{O}$. Similar to the graphs constructed above, each edge of +malicious (resp., -malicious) agreement graph represents one source maliciously endorses the other on the positive (resp., negative) claims of an object with a quantified endorsement degree, denoted as $\omega_{c_o}(s_1 \rightarrow s_2)$ (resp., $\tilde{\omega}_{c_o}(s_1 \rightarrow s_2)$), calculated by:

$$\omega_{c_o}(s_1 \rightarrow s_2) = \beta + (1 - \beta) \cdot \frac{|A_o(s_1, s_2)|}{|\mathcal{V}_{s_{2_o}}|} \cdot (1 - \prod_{v \in A_o(s_1, s_2)} \mathcal{C}_v) \cdot \mu(s_1, o) \tag{16}$$

$$\tilde{\omega}_{c_o}(s_1 \rightarrow s_2) = \beta + (1 - \beta) \cdot \frac{|\tilde{A}_o(s_1, s_2)|}{|\tilde{\mathcal{V}}_{s_{2_o}}|} \cdot (1 - \prod_{v \in \tilde{A}_o(s_1, s_2)} \mathcal{C}_{\tilde{v}}) \cdot \tilde{\mu}(s_1, o) \tag{17}$$

Both FPC random walk computation and normalization are conducted on each graph to obtain the dependence score for each source that provides positive (resp., negative) claims on an object o, denoted as $\mathcal{D}(s, o)$ (resp., $\tilde{\mathcal{D}}(s, o)$). We set the dependence score of the source with the highest visit probability in the +malicious agreement graph (resp., –malicious agreement graph) as pc_{max} (resp., nc_{max}). The computed dependence scores capture the following characteristics, all of which are consistent with our intuition:

- Vertices with more input edges should have a higher value of dependence score since those sources are maliciously endorsed by a larger number of sources. Such sources tend to act as collectors that copy values from several sources.
- The malicious endorsement. Similarly, a source that claims a value with a lower probability to be true (resp., false) in +malicious agreement graph should be more (resp., less) respected. Meanwhile, the malicious endorsement from a source on a value with a lower probability to be false (resp., true) in –malicious agreement graph, should be more (resp., less) respected.



(a) +Malicious Agreement Graph (b) –Malicious Agreement Graph

Figure 3 Sample ±malicious agreement graphs of four sources on book id : 9780072830613 in Table 2

- If a source shows more confidence in the claims (positive or negative) of the object, the endorsement from this source should be highlighted.

Example 3 Figure 3 shows sample ±malicious agreement graphs for book id : 9780072830613 (simply denoted as o) in the dataset described in Example 1[4]. Take the link from s_1 to s_2 in the sample +malicious agreement graph as an example, by applying (1), we get $A_o(s_1, s_2) = \{Stephen; James\}$, then $|A_o(s_1, s_2)| = 2$. By applying majority voting, we get the votes for values in $\mathcal{U}_o$ as {Stephen: 4, James: 2, Kate: 2, Merrill: 1}. Therefore, we initialize the confidence scores for the values as $\frac{4}{4} = 1$, $\frac{2}{4} = 0.5$, $\frac{2}{4} = 0.5$, $\frac{1}{4} = 0.25$. By substituting this result in (16), we obtain $\omega_{c_o}(s_1 \rightarrow s_2) = 0.1 + (1 - 0.1) \times \frac{2}{2} \times (1 - 1 \times 0.5) = 0.55$. In the same way, we obtain $\omega_{c_o}(s_1 \rightarrow s_3) = 0.1$, and $\omega_{c_o}(s_1 \rightarrow s_4) = 0.1$. Finally, the normalized weights of edges $s_1 \rightarrow s_2$, $s_1 \rightarrow s_3$, and $s_1 \rightarrow s_4$ of +malicious agreement graph are $\frac{0.55}{0.55+0.1+0.1} = 0.73$, $\frac{0.1}{0.55+0.1+0.1} = 0.135$, $\frac{0.1}{0.55+0.1+0.1} = 0.135$, respectively. After applying random walk computations, we obtain $\{\mathcal{D}(s_1, o) : 0.235, \mathcal{D}(s_2, o) : 0.245, \mathcal{D}(s_3, o) : 0.26, \mathcal{D}(s_4, o) : 0.26\}$, $\{\tilde{\mathcal{D}}(s_1, o) : 0.20, \tilde{\mathcal{D}}(s_2, o) : 0.25, \tilde{\mathcal{D}}(s_3, o) : 0.275, \tilde{\mathcal{D}}(s_4, o) : 0.275\}$. The results capture the relation patterns in the sample dataset: s_3 and s_4 are more likely to be copiers than other sources on either positive claims or negative claims for the specific book id : 9780072830613.

4.3 Object popularity quantification

Intuitively, popular objects tend to be covered by more sources, as sources tend to publish popular information to attract more audiences. Therefore, we quantify the popularity of an object, denoted as $\mathcal{P}_o$, by its occurrence rate. Specifically, each source casts a vote for the popularity of every object it covers, and each object collects votes from every source that

[4] We neglect the confidence scores of each source and omit the dependence score normalization step in this example.

claims values on it. We define the coverage of a source s, i.e., $Cov(s)$, as the percentage of its provided objects over $\mathcal{O}$. Formally, inspired by the idea from *term frequency-inverse document frequency* (i.e., TF-IDF) in information retrieval, we measure the popularity of each object by applying the following equations, which comprehensively incorporate the occurrence of the object and the coverage of each source that covers the object:

$$\mathcal{P}_o^u = \sum_{s \in \mathcal{S}_o} \frac{1}{Cov(s)} \tag{18}$$

where $\mathcal{P}_o^u$ is the un-normalized popularity of object o. $\mathcal{P}_o^u$ of all objects are then normalized as $\mathcal{P}_o$ to sum to 1.

The normalized popularity of each object captures the following two features, both of which are consistent with our intuition:

- Objects covered by more sources are more popular than those covered by fewer sources.
- Votes for the popularity of objects from the sources with lower coverage should be more respected than those from the sources with higher coverage.

4.4 Source confidence measurement

For a single-valued object, if a source claims a value for it, then the source certainly disclaims all the other values of the object. However, a straight application of the above mutual exclusion for a multi-valued object is unreasonable, as sources may not know the number of true values on the objects and therefore do not necessarily reject negative claims. To differentiate and quantify sources' confidence on their positive and negative claims, we incorporate loose mutual exclusion [36] into our model for source reliability calculation. The measurement approach of source confidence score is similar to the *Kappa coefficient* [20]. The main idea is to exclude the effect of random guess in determining the extent. In particular, the confidence score of s providing positive claims on o is calculated as:

$$\mu(s, o) = \frac{1}{|\mathcal{V}_{s_o}|} \cdot \left(1 - \frac{1}{|\mathcal{U}_o|}\right) \tag{19}$$

Meanwhile, the confidence score of s providing negative claims on o is calculated as:

$$\tilde{\mu}(s, o) = \frac{1}{|\tilde{\mathcal{V}}_{s_o}|} \cdot \frac{1}{|\mathcal{U}_o|} \tag{20}$$

The computed source confidence scores capture the following two features, which are consistent with our intuition:

- A cautious source, which only provides values that it is sure to be true and omits uncertain values, may claim partial true values of an object. Thus its confidence score on positive claims is relatively higher than that of the other sources. Similarly, its confidence score on negative claims is relatively lower than that of the other sources.
- For an audacious source, which tends to provide all potential values of an object, it may cover as many as possible values of an object, including false values. Therefore, its confidence score on positive claims is relatively lower than that of the other sources, while its confidence score on negative claims is relatively higher than that of the other sources.

Algorithm 1 The Algorithm of SmartVote

```
Input: objects of interest $\mathcal{O}$, a set of sources $\mathcal{S}$, and $\mathcal{V}_{s_o}$ the set of positive claims
       provided by each $s \in \mathcal{S}$ on each $o \in \mathcal{O}$.
Output: $\mathcal{V}_o{}^*$ identified truth for each $o \in \mathcal{O}$.
// Initialization phase
Initialize $\delta$, $\beta$, $\beta_{\mathcal{L}}$, $pp_{max}$, $np_{max}$, $pc_{max}$, $nc_{max}$
Initialize $\mathcal{C}_v, \mathcal{C}_{\tilde{v}}$ for each $v \in \mathcal{V}, o \in \mathcal{O}$
// Object popularity quantification
foreach $o \in \mathcal{O}$ do
    compute $\mathcal{P}_o$ by (18)
// Source confidence measurement
foreach $s \in \mathcal{S}$ do
    foreach $o \in \mathcal{O}$ do
        compute $\mu(s, o)$, $\tilde{\mu}(s, o)$ by (19), (20)
// Balancing long-tail phenomenon on source coverage
compute $\mathcal{L}(s_1, s_2)$ by (21)
// Iteration phase
repeat
    // Malicious agreement detection
    foreach $o \in \mathcal{O}$ do
        construct ±malicious agreement graphs by quantifying the weights of each
        edge by (16), (17)
        derive $\mathcal{D}(s, o)$, $\tilde{\mathcal{D}}(s, o)$ by applying random walk and normalization steps
    // ±SmartVote computation
    construct ±supportive agreement graphs by quantifying the weights of each edge
    by (10), (11),(12), (13)
    derive $\tau'(s)$, $\tilde{\tau}'(s)$ by applying random walk and normalization steps
    // Value confidence score computation
    foreach $v \in \mathcal{V}, o \in \mathcal{O}$ do
        compute $\mathcal{C}_v, \mathcal{C}_{\tilde{v}}$ by (14), (15)
until convergence;
return $\{(o, v) | v \in \mathcal{V} \wedge \mathcal{C}_v > \mathcal{C}_{\tilde{v}} \wedge v \in \mathcal{U}_o, o \in \mathcal{O}\}$
```

4.5 Balancing long-tail phenomenon on source coverage

In reality, many datasets show the long-tail phenomenon in their source coverage, i.e., very few sources provide extensive coverage for the objects of interest, while most of the sources only provide values for very few objects. Since identifying reliable sources is the key to determining value veracity and source reliability is typically estimated by the empirical probability of making correct claims, the accuracy of truth discovery and source reliability estimation depend on the coverage of the evaluated source. When sources cover numerous objects, we can conduct more accurate estimation of source reliability based on the sufficient evidence, leading to better truth discovery. However, due to the existence of the long-tail phenomenon, the majority of sources are "*small*" sources with very few claims. Source reliability estimation based on the limited evidence could be totally random. For example, consider the extreme case when most sources only cover one object. If the claimed values of one of these sources are correct and complete, the positive precision and negative precision of this source would both be one. On the other hand, if the claim is totally wrong, the positive precision would be zero.

To smooth the estimation for small sources, given an object, we consider three cases for the agreement between two sources: i) both sources sharing several common values; ii) both sources providing totally different values; iii) one source covering this object while the other source ignoring this object. To deal with the long-tail phenomenon, we assert that the agreement in the third case should not be zero. If a source does not cover an object, it does not indicate that this source votes against the values claimed by the other sources. In this section, our goal is to distinguish the third case from the second case. Formally, we use $\mathcal{L}(s_1, s_2)$ to represent a compensation for a link $s_1 \rightarrow s_2$, to re-balance the long-tail phenomenon on source coverage. In particular, for each object covered by s_2 but not covered by s_1, we approximately estimate the endorsement degree from s_1 to s_2 on this object according to (10) and (12). Each factor on the right side of $\sum$ in (21) corresponds to the factor in the same position of (10) and (12).

$$\mathcal{L}(s_1, s_2) = \begin{cases} \beta_{\mathcal{L}} \cdot \sum_{o \in \mathcal{O}_{s_2} - (\mathcal{O}_{s_1} \cap \mathcal{O}_{s_2})} \frac{1}{2} \cdot \frac{1}{2} \cdot \mathcal{P}_o \cdot \frac{1}{2} \cdot (\frac{1}{|\mathcal{U}_o|}(1 - \frac{1}{|\mathcal{U}_o|})); & \text{for } \mathcal{A}(s_1, s_2) \\ \beta_{\mathcal{L}} \cdot \sum_{o \in \mathcal{O}_{s_2} - (\mathcal{O}_{s_1} \cap \mathcal{O}_{s_2})} \frac{1}{2} \cdot \frac{1}{2} \cdot \mathcal{P}_o \cdot \frac{1}{2} \cdot \frac{1}{|\mathcal{U}_o|}^2; & \text{for } \tilde{\mathcal{A}}(s_1, s_2) \end{cases} \tag{21}$$

where $\beta_{\mathcal{L}}$ is an uncertainty factor of the compensation.

4.6 The algorithm

Algorithm 1 shows the whole procedure of SmartVote. The initialization phase initializes parameters with a priori values (line 1). The parameters to be initialized include iteration convergence threshold δ, smoothing factor β, uncertainty factor $\beta_{\mathcal{L}}$, positive precision pp_{max}, negative precision np_{max}, the two-sided dependence scores (pc_{max} and nc_{max}) of sources with the highest visit probabilities in both ±supportive agreement graphs and ±malicious agreement graphs. Besides, the confidence scores of each value v being true (denoted as $\mathcal{C}_v$) or being false (denoted as $\mathcal{C}_{\tilde{v}}$) are both initialized by adopting the majority voting in our experiments (alternative methods can be applied for this initialization). To start, we count the votes of each individual value of each object $o \in \mathcal{O}$. Then, we normalize those vote counts by dividing them by $|\mathcal{S}_o|$ to represent $\mathcal{C}_v$ for each value. $\mathcal{C}_{\tilde{v}}$ is initialized as $1 - \mathcal{C}_v$ (line 2). The *object popularity quantification* (lines 3-4), *source confidence measurement* (lines 5-7), and *long-tail phenomenon on source coverage balancing* (line 8) are calculated directly from the multi-sourced data outside the iteration. For each cycle of iteration, the algorithm does three parts of work: 1) recalculating the two-sided *source dependence scores* (lines 10-12); 2) calculating ±*SmartVote* (lines 13-14) of sources based on the two-sided *value confidence score*s; 3) computes value confidence scores (lines 15-16) based on ±SmartVote of sources. The algorithm is believed to converge when the difference of *cosine similarity* of ±SmartVote between two successive iterations turns smaller than a given threshold, δ (line 17).

The time complexity of the algorithm is $O(|\mathcal{O}||\mathcal{S}|^2 + |\mathcal{S}|^2 + |\mathcal{V}|)$. We believe the time complexity should not be an issue for the algorithm. First, there are many mature distributed computing tools that can be used for random walk computation to reduce the time complexity. For example, Apache Hama[5] is a framework for big data analytics, which uses the *Bulk*

[5] https://hama.apache.org/

Synchronous Parallel (BSP) computing model. It includes the *Graph* package for vertex-centric graph computations. Second, we can easily extend the *Vertex* class to create a class for realizing parallel random walk computation.

5 Experiments

5.1 Experimental setup

5.1.1 Real-world datasets

We used two real-world datasets in our experiments. Each object in both datasets may contain multiple true values. The *Book-Author dataset* [41] contains 33,971 book-author records crawled from *www.abebooks.com*. These records were collected from numerous book websites (i.e., sources). Each record represents a store's positive claims on the author name(s) of a book (i.e., object). We refined the dataset by removing the invalid and duplicated records, and excluding the records with only minor conflicts to make the problem more challenging—otherwise, even a straightforward method could yield competitive results. We finally obtained 13,659 distinctive claims, 624 websites providing values about author name(s) of 655 books. Each book has on average 3.1 authors. The ground truth provided by the original dataset was used as the gold standard. The *Parent-Children dataset* was extracted by focusing on the parent-children relation from the *Biography dataset* [30], which contains 11,099,730 records edited by different users about people's birth and death dates, names of their parents, children, and spouses on Wikipedia. We obtained 227,583 claims about 2,579 people's children information (i.e., objects) edited by 54,764 users (i.e., sources). We also further removed the duplicated and minor conflicting records for this dataset for more effective comparison. In the resulting dataset, each person has on average 2.48 children. We used the latest editing records as the gold standard.

5.1.2 Baseline methods

We compared SmartVote with three types of truth discovery methods as follows:

Existing MTD methods Based on our thorough analysis of the existing MTD methods (see Section 6), we chose the following three state-of-the-art MTD methods as the baselines and excluded other methods (e.g., [31, 35, 47]), which are already proved to perform worse than these three baselines [37]:

- *LTM (Latent Truth Model)* [46]: it applies a probabilistic graphical model to infer source reliability and value veracity.
- *MBM (Multi-truth Bayesian Model)* [36]: it incorporates source confidence and a finer-grained copy detection technique into a Bayesian model.
- *MTD-hrd* [38]: it is a model designed for *Multi-Truth Discovery*, which incorporates two implications, namely the calibration of imbalanced positive/negative claim distributions and the consideration of the implication of values' co-occurrence in the same claims, to improve the probabilistic approach.

STD methods Wang et al. [37] showed that all traditional STD methods achieve low accuracy in MTD scenarios when regarding a value set claimed by the same source as a single joint value. To adapt existing STD methods to MTD scenarios, we pre-processed the input datasets by conducting claim value separation. For example, for the record "s_2, 9780072830613, Stephen;James" in Table 2, we reformatted it into "s_2, 9780072830613, Stephen" and "s_2, 9780072830613, James". We thereby modified the STD methods to treat each value in a source's claimed value set on a given object individually, and determined the veracity of each individual value separately to accept multiple true values. We chose several representative methods but excluded those methods that are inapplicable to the MTD scenario for the comparison. For example, the methods proposed in [6] use the *number of false values* as prior knowledge, which is, however, impossible to be obtained in advance in the MTD scenario because we do not know the number of true values for each object. IATD (Influence-Aware Truth Discovery) [44] makes a number of variable distribution assumptions, which are unfeasible to adapt to the MTD scenarios as well. The method in [30] requires the normalization of the veracity scores of values, which is unfeasible for the MTD problem. The methods in [23, 45] focus on handling heterogeneous data, while our approach is designed specially for categorical data.

- *Voting*: this method regards a value as true if the proportion of the sources that claim this value exceeds a certain threshold.
- *Sums* [21], *Average-Log* [30]: these two methods are modified to incorporate mutual exclusion. They compute the total reliability of all sources that claim and disclaim a value separately. If the former is bigger than the latter, the value will be regarded as true.
- *TruthFinder* [41]: this method iteratively estimates *trustworthiness of source* and *confidence of fact* from each other by additionally considering the *influences between facts*.
- *2-Estimates* [17]: this method adopts mutual exclusion and recognizes a value as true if its truth probability exceeds 0.5.

Improved STD methods We improved the above STD methods by incorporating truth number prediction. In particular, for each method, we treated the values in the set of claimed values of each source individually and ran the original method to output source reliability and value confidence scores. Then, we computed $|\mathcal{V}_{s_o}|$ for each source on every object, based on which to predict the number of true values for each object:

$$P_o(n) = \sqrt[|\mathcal{S}_o|]{\prod_{|\mathcal{V}_{s_o}|=n, s\in\mathcal{S}_o} A(s) \cdot \prod_{|\mathcal{V}_{s_o}|\neq n, s\in\mathcal{S}_o} (1 - A(s))} \tag{22}$$

where $P_o(n)$ is the unnormalized probability[6] of the number of values of an object o to be n, and $A(s)$ is the reliability of s calculated by each method.

For each object, we chose the number with the highest probability (denoted as N) as the number of true values and output the top-N values instead of choosing the value set with the biggest confidence score as the outputs. Finally, we obtained five new methods, namely *Voting**[7], *Sums**, *Average-Log**, *TruthFinder**, and *2-Estimates**.

5.1.3 Parameter configuration

To ensure the fair comparison, we ran a series of experiments to determine the optimal parameter settings for each baseline method. We used the same criterion for all the iterative

[6] Such values are then normalized to represent probabilities.

[7] For Voting*, we predict the number of true values as the number with the highest vote counts.

methods to determine their convergence. For our approach, we simply used the default parameter settings for both datasets. In particular, we set $\beta_{\mathcal{L}}$ as 0.1, from our study on the impact of $\beta_{\mathcal{L}}$ on the performance of our approach (see Section 5.3.5). Intuitively, sources tend to provide values that they are sure to be true and omit uncertain values, while copiers are likely to copy those explicitly claimed values from other sources. Therefore, we set pp_{max} as 1, np_{max} as 0.9, pc_{max} as 1, and nc_{max} as 0.8.

5.1.4 Evaluation metrics

We implemented all the above methods in Python 3.4.0 and ran experiments on a 64-bit Windows 10 Pro. PC with an Intel Core i7-5600 processor and 16GB RAM. We ran each method multiple times (denoted as K, for our experiments, we set K as 10) to evaluate their average performance. In particular, we used two groups of evaluation metrics.

Traditional performance measurements *Precision* and *recall* are two commonly used performance measurements for evaluating the accuracy of truth discovery methods. We additionally used F_1 score to represent the overall accuracy. Execution time was measured for efficiency comparison.

Object popularity weighted performance measurements We introduced a new concept of object popularity to measure the importance of each object. Intuitively, objects with differed popularity have varied sizes of audience. The false values on an object with higher popularity would mislead people. Traditional performance metrics treat all the objects equally, thus cannot capture this implication. To measure the performance of truth discovery methods more precisely, we integrated object popularity as a weight into the calculation of precision, recall and F_1 score. We used the following three object popularity weighted performance metrics: i) *Weighted precision*, calculated as $\frac{1}{K}\sum_{k=1}^{K}\sum_{n=1}^{|\mathcal{O}|}\frac{|\mathcal{V}_o{}^{*(k)}\cap\mathcal{V}_o{}^{g}|}{|\mathcal{V}_o{}^{*(k)}|}\cdot\mathcal{P}_o$; ii) *Weighed recall*, calculated as $\frac{1}{K}\sum_{k=1}^{K}\sum_{n=1}^{|\mathcal{O}|}\frac{|\mathcal{V}_o{}^{*(k)}\cap\mathcal{V}_o{}^{g}|}{|\mathcal{V}_o{}^{g}|}\cdot\mathcal{P}_o$; and iii) *Weighted F_1 score*, which is the harmonious mean of weighted precision and weighted recall.

5.2 Comparative studies

Table 5 shows the performance of different methods on two real-world datasets in terms of accuracy and efficiency. For all the accuracy evaluation metrics except precision, *SmartVote* consistently achieved the highest value. Even in terms of precision, SmartVote still achieved the second-best performance on the Parent-Children dataset and the third-best performance on the Book-Author dataset. Among the four methods specially designed for the MTD problem, our approach is the most efficient as demonstrated by its lowest execution time. This is because *LTM* and *MTD-hrd* include complicated Bayesian inference over the complex probabilistic graphical model, and *MBM* conducts time-consuming copy detection, while our approach is based on a relatively simple graph model. All algorithms performed better on the Parent-Children dataset than on the Book-Author dataset. The possible reasons include the small scale, the poor quality of sources, and missing values (i.e., true values that are missed by all the data sources) of the Book-Author dataset, leading to insufficient evidence to support all correct values. The majority of the methods showed higher precision than recall, reflecting relatively high positive precision than the negative precision of most real-world sources.

Table 5 Comparison of different methods: the best and second best performance values are in bold

Method	Book-Author Dataset							Parent-Children Dataset						
	P	R	F1	WP	WR	WF1	T(s)	P	R	F1	WP	WR	WF1	T(s)
Voting	**0.84**	0.63	0.72	**0.83**	0.64	0.72	**0.07**	0.88	0.85	0.87	0.69	0.68	0.69	**0.56**
Sums	**0.84**	0.64	0.73	**0.83**	0.64	0.72	0.85	**0.90**	0.89	0.90	**0.88**	0.86	0.87	1.13
Avg-Log	**0.83**	0.60	0.70	**0.83**	0.64	0.72	0.61	**0.90**	0.89	0.89	**0.88**	0.86	0.87	**0.75**
TruthFinder	**0.84**	0.60	0.70	**0.83**	0.60	0.70	0.74	**0.90**	0.89	0.90	**0.88**	0.85	0.86	1.24
2-Estimates	0.81	0.70	0.75	0.80	0.68	0.74	0.38	**0.91**	0.89	0.90	**0.88**	0.86	0.87	1.34
Voting*	0.77	0.42	0.54	0.80	0.39	0.53	0.13	0.87	0.85	0.86	0.71	0.68	0.69	0.89
Sums*	**0.83**	0.24	0.38	**0.85**	0.21	0.34	0.99	0.86	0.88	0.87	0.67	0.84	0.75	1.45
Avg-Log*	0.74	0.49	0.59	0.80	0.53	0.64	**0.08**	0.89	0.87	0.88	0.77	0.82	0.79	0.92
TruthFinder*	0.70	0.71	0.70	0.75	**0.72**	0.73	0.99	0.85	0.91	0.88	0.69	0.88	0.77	1.16
2-Estimates*	**0.83**	0.24	0.38	0.81	0.21	0.34	0.79	0.86	0.89	0.87	0.66	0.83	0.74	1.47
LTM	0.82	0.65	0.73	0.82	0.62	0.71	0.98	0.87	0.90	0.88	0.86	0.89	0.87	0.99
MBM	**0.83**	**0.74**	**0.78**	0.82	0.71	**0.76**	0.67	**0.90**	**0.92**	**0.91**	0.87	**0.90**	**0.88**	2.17
MTD-hrd	**0.83**	0.58	0.68	**0.82**	0.59	0.69	0.72	**0.90**	0.90	**0.90**	0.87	0.89	**0.88**	1.37
SmartVote	0.81	**0.79**	**0.80**	**0.83**	**0.81**	**0.82**	0.45	**0.90**	**0.94**	**0.92**	**0.92**	**0.95**	**0.93**	0.92

Specifically, since *Voting* conducts truth discovery without iteration and the consideration of the quality of sources, it has relatively low accuracy, but on the other hand, it consumes the least execution time. The improved STD methods performed even worse than their original versions. This depicts that in reality, the majority of sources tend to be cautious and only provide values they are sure to be true. As a result, the predicted numbers of true values were generally smaller than the real ones, leading to lower precision and recall of the improved STD methods. Besides our approach, *2-Estimates* and *MBM* also performed better than the other methods in terms of both the traditional and weighted measurements. This can be attributed to their consideration of mutual exclusion. Though *LTM* and *MBM-hrd* also take this implication into consideration, they make strong assumptions on the prior distributions of latent variables. For this reason, once the dataset does not comply with the assumed distributions, it performs poorly. Without incorporating object popularity, *2-Estimates* and *MBM* showed lower quality Compared with *MBM*, which showed second-best performance, *SmartVote* not only includes object popularity but also globally models two types of source relations and pays attention to the ubiquitous long-tail phenomenon on source coverage. Overall, *SmartVote* showed the best accuracy performance.

5.3 Impact of different concerns

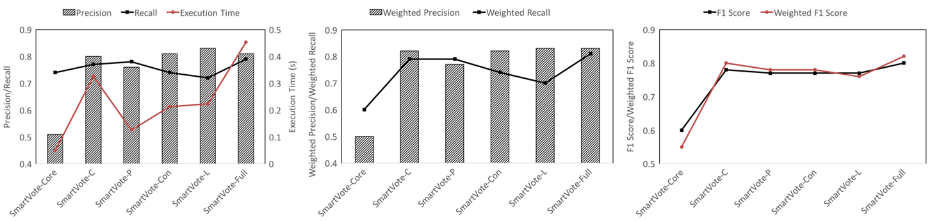
The compound effect of different technical components contributes to the performance of SmartVote. To evaluate the impact of different concerns, we implemented five variants of SmartVote:

- *SmartVote-Core*: A variant of SmartVote without incorporating the four implications.
- *SmartVote-C*: A version of SmartVote-Core that adopts the malicious agreement detection.
- *SmartVote-P*: A version of SmartVote-Core that adopts the object popularity quantification.
- *SmartVote-Con*: A version of SmartVote-Core that incorporates the loose mutual exclusion.
- *SmartVote-L*: A version of SmartVote-Core that considers the long-tail phenomenon on source coverage.

Figure 4 reports the performance comparison of different variants of SmartVote on the Book-Author dataset. The experimental results on the Parent-Children dataset show the similar insights. By incorporating each individual component into SmartVote-Core, our approach showed consistently better performance in terms of accuracy with only slight increase in the execution time. The full version of SmartVote led to the best result. We studied the impact of each technical component on our approach and report the findings in the following sections.

5.3.1 SourceVote

To validate the feasibility of modeling source reliability by quantifying the two-sided inter-source agreements, we implemented *SourceVote* to estimate source reliability as ±SourceVote for the two real-world datasets. We used these results to initialize the parameters regarding source reliability of the aforementioned baseline methods, including *Sums, Average-Log, TruthFinder, 2-Estimates, LTM*, and *MBM*. Note that we did not apply *SourceVote* to *Voting*, because *Voting* assumes all sources are equally reliable.



(a) Precision, Recall & T(s) (b) Weighted Precision & Recall (c) F1 Score & Weighted F1 Score

Figure 4 Performance comparison of different variants of SmartVote

Figure 5 describes the performance comparison of the SourceVote initialized methods with their original versions in terms of precision, recall, and execution time on the Book-Author dataset. We omit the results on the Parent-Children dataset as it led to similar conclusions. The results show that initializing source reliability by applying *SourceVote* almost led all methods to better performance, indicated by higher precision and recall, lower execution time. This reflects that the source reliability evaluated by *SourceVote* is more accurate than the widely applied default value of 0.8. With precise initialization, all methods show a faster convergence speed. Especially, the precision and recall of *MBM* stayed stable. This indicates that *MBM* is insensitive to the initial assumptions of source quality. However, the execution time of *MBM* was reduced dramatically due to the precise source quality inputs. The execution time of *LTM* increased because the number of iterations was fixed to 1,001 to ensure algorithm convergence and avoid performance fluctuations (as suggested in [34]). The increased execution time came from the time spent on *SourceVote*.

5.3.2 Malicious agreement

To validate the fact that two types of source relations, namely source supportive relations and copying relations (which corresponds to source supportive agreements and malicious

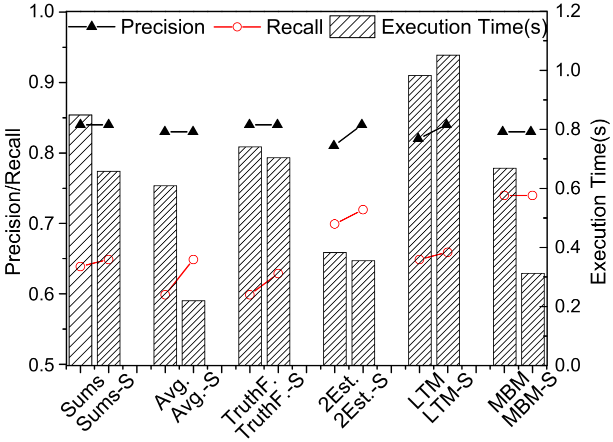


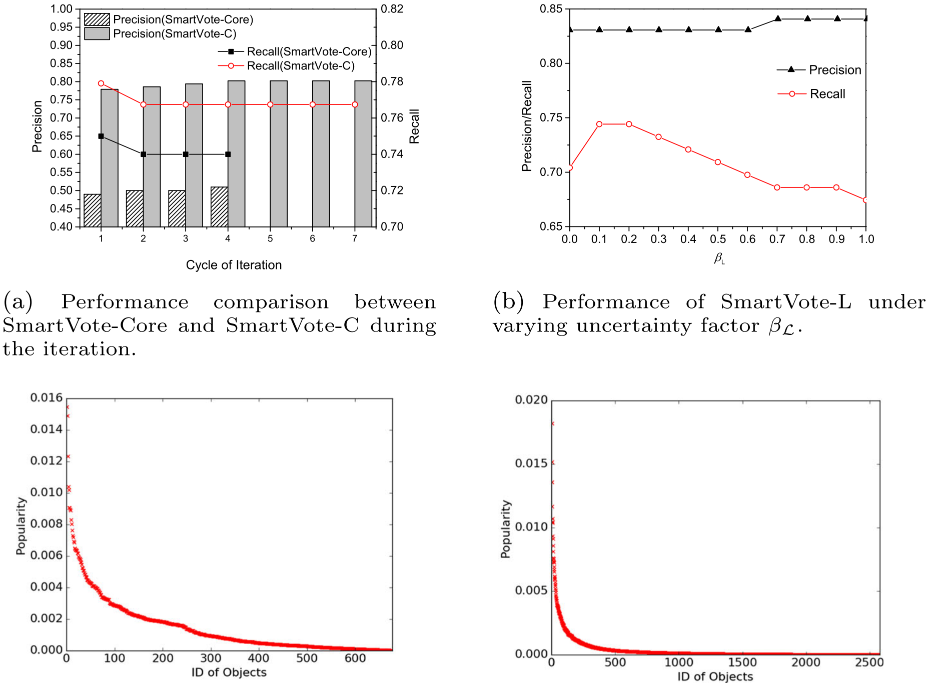
Figure 5 Comparison between the original versions of representative existing truth discovery methods and the versions that apply SourceVote for precise source reliability initialization. The latter versions are marked by suffix "-s"

agreements), widely exist in real-world datasets, we conducted an analysis on the Book-Author dataset. The analysis results on the Parent-Children dataset show the same features. Take the book (id : 9780072223194) as an example, the ground truth shows that “Dennis Suhanovs” is its true author. From the dataset, we can see that the false value “Laura Robinson” is claimed by three sources including “The Recycled Book Shop”, “Blackwell Online”, and “textbookxdotcom”. This phenomenon indicates the possible malicious agreements (or copying relations) among the three sources. Meanwhile, the true value “Dennis Suhanovs” is claimed by the majority of sources, which indicates supportive agreements (or supportive relations) among those sources. Overall, among all the books covered by the ground truth, we found only 11.76%, on which no source shows malicious agreement with others. For this small portion of objects, sources make unique false claims other than common false claims. Meanwhile, there is only one book, on which no source claims the true values, and therefore no supportive relation exists among those sources on this book. Based on this analysis, we conclude that the supportive relations and copying relations are ubiquitous in real-world datasets. That is why we should consider those relations when pursuing better truth discovery.

By incorporating malicious agreement detection component, both precision, recall and F_1 score of SmartVote-C are higher than SmartVote-Core, as shown in Figure 4. Interestingly, when we leveraged weighted metrics to evaluate SmartVote-C, the algorithm even showed better results than by using traditional metrics. The results report the wide existence of copying relations in real-world datasets. Neglecting these relations would lead to the result of overestimating the reliability of copiers and impair the performance of truth discovery methods. Among other components, malicious agreement detection is the most time-consuming, as we need to compute the dependence score of each source on each object iteratively from the confidence scores of the claimed values of each object, and calculate the reliability of each source iteratively from the independence score of each source and the confidence score of each value. However, when comparing with the performance improvement introduced by incorporating this component, this additional amount of time cost can be justified. To study the effect of this component in depth, we compared the performance of SmartVote-Core and SmartVote-C in terms of precision and recall for each cycle of iteration, as shown in Figure 6a. Although SmartVote-C took longer time to converge, i.e., 7 rounds of iteration (SmartVote-Core only required 4 rounds of iteration), it consistently achieved higher performance in each round of iteration.

5.3.3 Object popularity

By considering the different popularity of objects, SmartVote-P performed better in terms of accuracy with nearly no extra execution time cost. This is because more sources provide claims on popular objects, and more evidence can be obtained to model the endorsement among sources. Therefore, when computing source reliability, assigning more weights to the popular objects would lead to better truth discovery. In addition, object popularity is calculated directly from the multi-sourced data. Since this calculation is outside of the iteration, it can be conducted effectively under linear execution time. Another observation was that SmartVote-P achieved higher weighted accuracy. This is consistent with our expectation because source reliability evaluation relies on the claims provided on popular objects. By differentiating the popularity of objects, our approach obtained more precise results. By ranking objects in the Book-Author dataset and the Parent-Children dataset, respectively, in a descending order of their popularity degrees, we draw scatter diagrams as shown in Figure 6c and d, where each point depicts an object with the corresponding popularity



(a) Performance comparison between SmartVote-Core and SmartVote-C during the iteration.

(b) Performance of SmartVote-L under varying uncertainty factor $\beta_{\mathcal{L}}$.

(c) Distribution of object popularity in Book-Author dataset

(d) Distribution of object popularity in Parent-Children dataset

Figure 6 Impact of different concerns

degree (totally, there are 677 objects in the Book-Author dataset and 2,579 objects in the Parent-Children dataset). We observed that in both scatter diagrams, the points with very high popularity degrees are quite sparse, indicating that only very few objects are more popular than the majority.

To further validate SmartVote, we compared with MBM (the best baseline method) on the top-20 popular objects in the ground truth of the Book-Author dataset. SmartVote returned false values on 2 objects (Book id : 9780072499544 and Book id : 9780071362856) while MBM made mistakes on 4 objects (Book id : 9780028056005, Book id : 9780072499544, Book id : 9780071362856, and Book id : 9780072843996). This demonstrated that SmartVote had better accuracy on the more popular objects. SmartVote and MBM both returned false values on Book id : 9780072499544 and Book id : 9780071362856 because some authors are neglected by all the sources.

5.3.4 Source confidence

We can see from Figure 4 that SmartVote-Con performed better than SmartVote-Core in terms of precision and F_1 score while keeping the recall unchanged. The reason for the better performance of SmartVote-Con is that in real-world datasets, the total number of distinct values of an object provided by all the sources is generally much larger than the number of positive claims of a specific source. Thus, sources normally make more negative claims than positive claims and show different confidence for these two types of claims. Neglecting this type of differences and strictly conducting the mutual exclusion would certainly increase the false negatives of the truth discovery methods. On the other hand, re-balancing and quantifying the confidence scores of sources for their positive claims and negative claims according to their distributions in the datasets make our approach more aligned with the reality.

5.3.5 Long-tail phenomenon on source coverage

Incorporating the balancing long-tail phenomenon on source coverage component dramatically increased the precision of SmartVote-Core with only a slight decrease in the recall, resulting in a higher value of F_1 score. In reality, different sources often cover different objects. For the case that a source covers an object that is ignored by another source, the source reliability will be under-estimated if we directly consider there is no agreement between the two sources. On the other hand, the source reliability may be over-estimated, if we use the average endorsement degree between the two sources on the commonly covered objects to measure the overall endorsement degree between the sources. Our approach tackles this issue by modeling the claim distributions, and experimental results validate the effectiveness of our approach. We also investigated the performance of our approach by tuning the values of the uncertainty factor $\beta_{\mathcal{L}}$ from 0 to 1 (as shown in Figure 6b). We found that the precision stayed stable for varying values of $\beta_{\mathcal{L}}$, while the recall peaked at the points when $\beta_{\mathcal{L}}$ equals to 0.1 and 0.2.

5.3.6 Convergence discussion

In this experiment, different variants of SmartVote all converged on both Book-Author and Parent-Children datasets. As shown in Table 5, the execution time of SmartVote is the shortest among all the compared MTD methods, indicating that SmartVote converged at a quicker rate than those methods. Take SmartVote-Core and SmartVote-C as examples, we can see from Figure 6a that SmartVote-Core converged at the 4th round of iteration, while SmartVote-C converged at a slower rate (required 7 rounds of iteration). According to a recent survey [26], the open question on "*will the existing truth discovery methods promise convergence or not*" still remains unsolved. The study on the convergence rate and how to bound the errors of converged results of the methods could be an interesting future research direction. We leave the theoretical analysis of the convergence of our approach as future work.

6 Related work

Due to the significance of the veracity to *Big Data*, truth discovery has been a hot topic and studied actively for last few years in the database community [12, 13]. Aiming at resolving the conflicts among the multi-sourced data, and determining the underlying true values, significant research efforts have been conducted and many methods have been proposed for truth discovery in various application scenarios (see [24, 26, 34] for surveys). The *primitive methods* are typically *rule-based*, such as: i) regarding the latest edited values as true; ii) conducting *majority voting* (for categorical data), i.e., predicting the values with the highest number of occurrences as the truths; iii) naively taking the *mean/minimum/maximum* as the true values (for continuous data). These methods focus on improving the efficiency in database queries [2, 3, 8, 29], but they show low accuracy for cases that many sources provide low-quality data, due to the fact that the sources may not be equally reliable.

Yin et al. [41] first formulate the truth discovery problem in 2008. Since then, many advanced solutions have been proposed by applying unsupervised or semi-supervised learning techniques while additionally taking various implications of multi-sourced data into consideration. According to the models they adopted, we classify the advanced methods into five categories[8].

The first category is *link* based methods [21, 30] that conduct random walks on the bipartite graph between sources and values of objects. They measure source authority based on their links to the claimed values and estimate source reliability and value correctness based on the bipartite graph. *Iterative* methods [17, 30, 41] iteratively calculate value veracity and source reliability from each other until certain convergence condition is met. *Bayesian point estimation* methods [6, 36] adopt *Bayesian analysis* to compute the maximum a posteriori probability or *MAP* value for each object. *Probabilistic graphical model* based methods [38, 45, 46] apply probabilistic graphical models to jointly reason about source trustworthiness and value correctness. Finally, *optimization* based methods [22, 23] formulate the truth discovery problem as an optimization problem. Recently, Popat et al. [32] propose an approach for early detection of emerging textual claims. This is a very interesting direction of truth discovery, but out of the scope of our paper.

Since source reliability is the key to determining value veracity and existing truth discovery methods generally require source reliability initialization to launch their algorithm, more precise source reliability initialization is much in demand. Recent work adopts an external trustful source [11], a subset of labeled data [9, 27, 42], or the similarity among sources [43] as prior knowledge to initialize or help initialize the source reliability. To the best of our knowledge, SourceVote is the very first few to help source reliability initialization nearly without any prior knowledge.

To realize better truth discovery, the long-tail phenomenon on source coverage of multi-sourced data has been empirically investigated on four real-world datasets by Li et al. in [22]. The observations of sources' authority features and sources' copying relations have been presented in [21] and [5–7, 25, 44]. The long-tail phenomenon on objects has also been studied by Xiao et al. [39] to conduct confidence interval estimation. However, all those implications have been incorporated into different models under the single-value assumption.

Despite active research in the field, multi-truth discovery (MTD) is rarely studied by the previous work. LTM (Latent Truth Model) [46], a probabilistic graphical model-based method, is the first solution to MTD. In this work, Zhao et al. measure two types of errors (false positive and false negative) by modeling two different aspects of source reliability (*specificity* and *sensitivity*) in a generative process. The disadvantage is that LTM makes strong assumptions about prior distributions for nine latent variables, rendering the model inhibitive and intractable to incorporating various implications to improve its performance. Pochampally et al. [31] study various correlations among sources by taking information extractors into consideration, but their application scenario is different from ours. The

[8] Note that there are overlaps among those categories. For example, Investment belongs to both Web-link based methods and iterative methods.

experiments show that their basic model without considering source correlations sometimes performs worse than LTM. In our experiments, SmartVote constantly achieves better results than LTM. To rebalance the distributions of positive claims and negative claims and to incorporate the implication of values' co-occurrence in the same claims, Wang et al. [38] propose a probabilistic model that takes multi-valued objects into consideration. However, this method also requires initialization of multiple parameters, such as prior true or false count of each object, and prior false positive or true negative count of each source.

Waguih et al. [34] conclude with extensive experiments that these probabilistic graphical model-based methods cannot scale well. Zhi et al. [47] also consider the mutual exclusion between sources' positive claims and negative claims, but they model the silence rate of sources to tackle the possible non-truth objects rather than multi-valued objects. To relax unnecessary assumptions, Wang et al. [36] analyze the unique features of MTD and propose MBM (Multi-truth Bayesian Model), which incorporates source confidence and finer-grained copy detection techniques in a Bayesian framework. However, they assume that false information is copied from sources and correct information is provided independently by sources. Recently, Wang et al. [37] design three models (i.e., the *byproduct* model, the *joint* model and the *synthesis* model) for enhancing existing truth discovery methods. Their experiments show that those models are effective in improving the accuracy of multi-truth discovery using existing truth discovery methods. However, LTM and MBM still perform better than those enhanced methods. Wan et al. [35] propose an uncertainty-aware approach for the real-world cases where the number of true values is unknown. However, they deal with continuous data rather than categorical data.

Different from those methods, SmartVote is a graph-based method [15], which incorporates four types of implications. In particular, SmartVote has four features: i) it is the first to take the impact of object popularity on source reliability into consideration; ii) instead of assuming independence of sources (in LTM) or independent copying relations among sources (in MBM), SmartVote globally models copying relations by constructing graphs of all sources that provide values on a specific object; iii) different from previous copy detection approaches, including MBM and other methods [5–7, 31], which only consider the copying relations among sources, SmartVote not only punishes *malicious copiers* that make the same faults with other sources but also defines a new source relation, namely *supportive relation*, to punish that sources that support each other by providing the same true values; and iv) as the long-tail phenomenon on source coverage is common in reality, SmartVote additionally deals with this significant issue by avoiding sources with few claims from being assigned high reliability.

A recent survey [26] shows that most of truth discovery methods start with uniform weights among all sources, and as a result, the performance of truth discovery may rely on the majority. Therefore, it is safe to say that most existing truth discovery methods may fail when the majority of sources are of low-quality. Typically, for the special circumstances where most of sources are malicious copiers or of low-quality, prior knowledge regarding value veracity from the Web content is required; otherwise, there would be no clue to reason about the truth. A promising work that could relieve this issue is proposed by Yu et al. [43], who develop a novel unsupervised multi-dimensional truth-finding framework to explore the wisdom of minority. They leverage the strengths of both Natural Language Processing work and Data Mining work. We leave the research on multi truth discovery under such circumstances as our future work.

7 Conclusion

In this paper, we focus on the problem of discovering true values for multi-valued objects (i.e., multi truth discovery problem or MTD). We first propose an approach, called *SourceVote*, to model two aspects of source reliability (i.e., positive precision and negative precision) by quantifying the two-sided inter-source agreements. SourceVote can be utilized to initialize the existing truth discovery methods. To further improve the accuracy of source reliability estimation and predict truths for multi-valued objects, we further propose a full-fledged graph-based model, *SmartVote*. This model incorporates four implications including source relations (i.e., *supportive relations* and *copying relations*), *object popularity*, *loose mutual exclusion*, and *long-tail phenomenon on source coverage*. We construct ±supportive agreement graphs to model the endorsement among sources on their positive and negative claims, from which to derive improved evaluations of two-sided source reliability. We then capture copying relations among sources by constructing the ±malicious agreement graphs based on the consideration that sources sharing the same false values are more likely to be dependent. We further consider the popularity of objects and develop techniques to quantify object popularity based on object occurrences and source coverage. Finally, we apply source confidence scores to differentiate the extent to what a source believes its positive claims and negative claims. For the ubiquitous long-tail phenomenon on source coverage, we introduce smoothing weights to the ±supportive agreement graphs to avoid the reliability of small sources from being over- or under-estimated. Experimental results show that our approach outperforms the state-of-the-art truth discovery methods on two real-world large datasets. Inspired by a recent research effort by Xiao et al. [40], one of our future research efforts will focus on convergence analysis of our truth discovery approach and investigate the theoretical guarantee of the results of our approach.

Acknowledgements Quan Z. Sheng's research has been partially supported by Australian Research Council (ARC) Future Fellowship FT140101247 and Discovery Project Grant DP180102378. Dianhui Chu's research has been partially supported by National Science Foundation of China (NSFC, No 61772159). The authors would like to thank the anonymous reviewers for their valuable feedback on this work.

References

1. Benslimane, D. et al.: The uncertain Web: concepts, challenges, and current solutions. ACM Transactions on Internet Technology (TOIT) **16**(1), 1 (2015)
2. Bleiholder, J., Naumann, F.: Conflict handling strategies in an integrated information system. In: Proceedings of the Intelligence Workshop on Information Integration on the Web (IIWeb) (2006)
3. Bleiholder, J., Naumann, F.: Data fusion. ACM Computing Surveys (CSUR) **41**(1), 1–41 (2009)
4. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Comput. Netw. ISDN Syst. **30**(1–7), 107–117 (1998)
5. Dong, X.L., Berti-Equille, L., Hu, Y., Srivastava, D.: Global detection of complex copying relationships between sources. Proc. VLDB Endowment **3**(1-2), 1358–1369 (2010)
6. Dong, X.L., Berti-Equille, L., Srivastava, D.: Integrating conflicting data: the role of source dependence. Proc. VLDB Endowment **2**(1), 550–561 (2009)

7. Dong, X.L., Berti-Equille, L., Srivastava, D.: Truth discovery and copying detection in a dynamic world. Proc. VLDB Endowment **2**(1), 562–573 (2009)
8. Dong, X.L., Naumann, F.: Data fusion: resolving data conflicts for integration. Proc. VLDB Endowment **2**(2), 1654–1655 (2009)
9. Dong, X.L., Saha, B., Srivastava, D.: Less is more: selecting sources wisely for integration. Proc. VLDB Endowment **6**(2), 37–48 (2012)
10. Dong, X.L., et al.: From data fusion to knowledge fusion. Proc. VLDB Endowment **7**(10), 881–892 (2014)
11. Dong, X.L., et al.: Knowledge vault: a Web-scale approach to probabilistic knowledge fusion. In: Proceedings of the ACM SIGKDD Intelligence Conference on Knowledge Discovery and Data Mining, pp. 601–610 (2014)
12. Fan, W.: Data quality: theory and practice. In: Web-Age Information Management, pp. 1–16 (2012)
13. Fan, W., et al.: Data quality problems beyond consistency and duduplication. In: Search of Elegance in the Theory and Practice of Computation, pp. 237–249 (2013)
14. Fang, X.S.: Generating actionable knowledge from big data. In: Proceedings of the 2015 SIGMOD Phd Symposium (SIGMOD), pp. 3–8 (2015)
15. Fang, X.S., Sheng, Q.Z., Wang, X., Ngu, A.H.: Value veracity estimation for multi-truth ojbects via a graph-based approach. In: Proceedings of the Intelligence World Wide Web Conference (WWW), pp. 217–226 (2017)
16. Fang, X.S., Wang, X., Sheng, Q.Z.: Ontology augmentation via attribute extraction from multiple types of sources. In: Proceedings of the 26Th Australasian Database Conference (ADC), pp. 16–27 (2015)
17. Galland, A., et al.: Corroborating information from disagreeing views. In: Proceedings of the ACM Intelligence Conference on Web Search and Data Mining (WSDM), pp. 131–140 (2010)
18. Gao, J., Li, Q., Zhao, B., Fan, W., Han, J.: Truth discovery and crowdsourcing aggregation: a unified perspective. Proc. VLDB Endowment **8**(12), 2048–2049 (2015)
19. Gleich, D.F. et al.: Tracking the random surfer: empirically measured teleportation parameters in pagerank. In: Proceedings of the Intelligence World Wide Web Conference (WWW), pp. 381–390 (2010)
20. Gwet, K.L.: Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters. Adv. Anal. LLC **4**, 57–64 (2014)
21. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632 (1999)
22. Li, Q., et al.: A confidence-aware approach for truth discovery on long-tail data. Proc. VLDB Endowment **8**(4), 425–436 (2014)
23. Li, Q., et al.: Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In: Proceedings ACM SIGMOD Intelligence Conference on Management of Data, pp. 1187–1198 (2014)
24. Li, X., Dong, X.L., Lyons, K., Meng, W., Srivastava, D.: Truth finding on the deep web: is the problem solved? Proc. VLDB Endowment **6**(2), 97–108 (2012)
25. Li, X., et al.: Scaling up copy detection. In: IEEE Intelligence Conference on Data Engineering (ICDE), pp. 89–100 (2015)
26. Li, Y., et al.: A survey on truth discovery. ACM SIGKDD Explor. Newslett. **17**(2), 1–16 (2016)
27. Liu, X., et al.: Online data fusion. Proc. VLDB Endowment **4**(11), 932–943 (2011)
28. Mukherjee, S., et al.: People on drugs: credibility of user statements in health communities. In: ACM SIGKDD Intelligence Conference on Knowledge Discovery and Data Mining, pp. 65–74 (2014)
29. Naumann, F., et al.: Data fusion in three steps: resolving schema, tuple, and value inconsistencies. IEEE Data Eng. Bull. **29**(2), 21–31 (2006)
30. Pasternack, J., Roth, D.: Knowing what to believe (when you already know something). In: Proceedings of Intelligent Conference on Computational Linguistics (COLING), pp. 877–885 (2010)
31. Pochampally, R. et al.: Fusing data with correlations. In: Proceedings of the ACM SIGMOD Intelligent Conference on Management of Data, pp. 433–444 (2014)
32. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G.: Where the truth lies: explaining the credibility of emerging claims on the Web and social media. In: Proceedings Intelligent World Wide Web Conference (WWW), pp. 1003–1012 (2017)
33. Rozenshtein, P., Anagnostopoulos, A., Gionis, A., Tatti, N.: Event detection in activity networks. In: Proceedings of the ACM SIGKDD Intelligent Conference on Knowledge Discovery and Data Mining, pp. 1176–1185 (2014)
34. Waguih, D.A., Berti-Equille, L.: Truth discovery algorithms: an experimental evaluation. arXiv:1409.6428 (2014)
35. Wan, M. et al.: From truth discovery to trustworthy opinion discovery: an uncertainty-aware quantitative modeling approach. In: Proceedings of the ACM SIGKDD Intelligent Conference on Knowledge Discovery and Data Mining, pp. 1885–1894 (2016)

36. Wang, X., et al.: An integrated Bayesian approach for effective multi-truth discovery. In: Proceedings the 24th ACM Intelligent Conference on Information and Knowledge Management (CIKM), pp. 493–502 (2015)
37. Wang, X., et al: Empowering truth discovery with multi-truth prediction. In: Proceedings the 25th ACM Intelligent Conference on Information and Knowledge Management (CIKM), pp. 881–890 (2016)
38. Wang, X., et al.: Truth discovery via exploiting implications from multi-source data. In: Proceedings the 25th ACM Intelligent Conference on Information and Knowledge Management (CIKM), pp. 861–870 (2016)
39. Xiao, H., Gao, J., Li, Q., Ma, F., Su, L., Feng, Y., Zhang, A.: Towards confidence in the truth: a bootstrapping based truth discovery approach. In: Proceedings ACM SIGKDD Intelligent Conference on Knowledge Discovery and Data Mining, pp. 1935–1944 (2016)
40. Xiao, H., Gao, J., Wang, Z., Wang, S., Su, L., Liu, H.: A truth discovery approach with theoretical guarantee. In: Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016, pp. 1925–1934 (2016)
41. Yin, X., Han, J., Yu, P.S.: Truth discovery with multiple conflicting information providers on the Web. IEEE Transactions on Knowledge and Data Engineering (TKDE) **20**(6), 796–808 (2008)
42. Yin, X., et al.: Semi-supervised truth discovery. In: Proceedings Intelligent World Wide Web Conference (WWW), pp. 217–226 (2011)
43. Yu, D. et al.: The wisdom of minority: unsupervised slot filling validation based on multi-dimensional truth-finding. In: Proceedings Intelligent Conference on Computational Linguistics (COLING), pp. 1567–1578 (2014)
44. Zhang, H., Li, Q., Ma, F., Xiao, H., Li, Y., Gao, J., Su, L.: Influence-aware truth discovery. In: Proceedings the 25th ACM Intelligent Conference on Information and Knowledge Management (CIKM), pp. 851–860 (2016)
45. Zhao, B., Han, J.: A probabilistic model for estimating real-valued truth from conflicting sources. In: Proceedings of the Intelligent Workshop on Quality in Databases (QDB), Coheld with VLDB (2012)
46. Zhao, B., Rubinstein, B.I., Gemmell, J., Han, J.: A bayesian approach to discovering truth from conflicting sources for data integration. Proc. VLDB Endowment **5**(6), 550–561 (2012)
47. Zhi, S., Zhao, B., Tong, W., Gao, J., Yu, D., Ji, H., Han, J.: Modeling truth existence in truth discovery. In: Proceedings ACM SIGKDD Intelligent Conference on Knowledge Discovery and Data Mining, pp. 1543–1552 (2015)

Affiliations

Xiu Susie Fang[1] · Quan Z. Sheng[1] · Xianzhi Wang[2] · Dianhui Chu[3] · Anne H. H. Ngu[4]

Xianzhi Wang
sandyawang@gmail.com

Dianhui Chu
chudh@hit.edu.cn

Anne H. H. Ngu
angu@txstate.edu

[1] Department of Computing, Macquarie University, Sydney, NSW 2109, Australia

[2] Living Analytics Research Center, School of Information Systems, Singapore Management University, 178902 Singapore, Singapore

[3] School of Computer Science and Technology, Harbin Institute of Technology (Weihai), Weihai, China

[4] Department of Computer Science, Texas State University, San Marcos, TX 78666, USA