Truth Discovery via Exploiting Implications from Multi-Source Data

Xianzhi Wang[1], Quan Z. Sheng[2], Lina Yao[1], Xue Li[3],
Xiu Susie Fang[2], Xiaofei Xu[4], and Boualem Benatallah[1]
[1]University of New South Wales, Sydney, NSW 3052, Australia
[2]University of Adelaide, Adelaide, SA 5005, Australia
[3]University of Queensland, Brisbane, QLD 4072, Australia
[4]Harbin Institute of Technology, Harbin, 150001, China
{xianzhi.wang, lina.yao, boualem.benatallah}@unsw.edu.au, {michael.sheng, xiu.fang}@adelaide.edu.au, xueli@itee.uq.edu.au, xiaofei@hit.edu.cn

ABSTRACT

Data veracity is a grand challenge for various tasks on the Web. Since the web data sources are inherently unreliable and may provide conflicting information about the same real-world entities, truth discovery is emerging as a countermeasure of resolving the conflicts by discovering the truth, which conforms to the reality, from the multi-source data. A major challenge related to truth discovery is that different data items may have varying numbers of true values (or multi-truth), which counters the assumption of existing truth discovery methods that each data item should have exactly one true value. In this paper, we address this challenge by exploiting and leveraging the implications from multi-source data. In particular, we exploit three types of implications, namely *the implicit negative claims*, *the distribution of positive/negative claims*, and *the co-occurrence of values in sources' claims*, to facilitate multi-truth discovery. We propose a probabilistic approach with improvement measures that incorporate the three implications in all stages of truth discovery process. In particular, incorporating the negative claims enables multi-truth discovery, considering the distribution of positive/negative claims relieves truth discovery from the impact of sources' behavioral features in the specific datasets, and considering values' co-occurrence relationship compensates the information lost from evaluating each value in the same claims individually. Experimental results on three real-world datasets demonstrate the effectiveness of our approach.

Keywords

Truth discovery; multiple true values; probabilistic model; imbalanced claims

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA

DOI: http://dx.doi.org/10.1145/2983323.2983791

1. INTRODUCTION

The World Wide Web (WWW) has transformed radically in recent years as a platform for collecting, storing, processing, managing, and querying the Big Data [2]. Besides the exploding amount of data, the type and number of data sources on the Web have increased enormously. Each day, around 2.5 quintillion bytes of data are created from various sources, such as sensors in the Web of Things (WoT) applications, posts in social networking websites, and transaction records in e-Commerce systems [3]. A common requirement of these applications is to integrate and exploit useful information from the multi-source data. The sequent issue is that the data sources in an open environment are inherently unreliable—they may provide incomplete, out-of-date, or even erroneous data [2]. For example, sensors in wild fields may produce inaccurate readings due to hardware limitations or device malfunction, weather websites may publish out-of-date weather information due to delayed updates, and workers in crowdsourcing systems may assign different labels to the same picture as a result of their varying expertise and biases. Moreover, it is common in e-Commerce environments that sellers attract customers by posting unreal low prices. Consequently, given a real-world entity, different data sources are liable to provide conflicting data. It therefore becomes important to discover the truth from the multi-source data to resolve the conflicts.

The main challenge regarding truth discovery is that, given a specific data item, the number of true values is generally unknown. For example, many online bookstores, such as textbooksNow.com and textbookx.com list *Miles J. Murdocca* as the only author of the book "Principles of Computer Architecture", while other stores, such as A1Books and ActiniaBookstores, post *J Miles Murdocca* and *Heuring P Vincent* as co-authors of the same book. Given the varying numbers of authors in the conflicting records, users may find it difficult to determine the correct authors of the book. Although many truth discovery methods are been proposed previously, they are mostly designed to discover a single true value for each data item; so given a data item that possesses multiple true values, they simply regard the values claimed by each source as a joint single value and determine the truthfulness of these values together. This is unreasonable since the sets of values claimed by different sources are generally correlated and should not be evaluated independently. Neglect of this hint could greatly degrade the truth discov-

ery accuracy. Take the above bookstores for example, the values claimed by textbookx.com and A1Books are not totally different. By claiming *J Miles Murdocca* and *Heuring P Vincent* to be the co-authors of the book, A1Books implicates that *J Miles Murdocca* is a correct author; similarly, by claiming *Miles J. Murdocca* as the author of the book, textbookx.com, in turn, partially supports A1Books's claim of the two authors.

In this paper, we study the problem of discovering varying numbers of true values of different data items from multi-source data, or the multi-truth discovery problem (MTD). We propose a probabilistic approach with improvement measures that leverages the implications of the multi-source data throughout all stages to facilitate multi-truth discovery. In a nutshell, we make the following contributions:

- We formally define the MTD and propose a probabilistic approach, which separates the consideration of each value in the same claims and takes into account the impact of both positive claims and negative claims, which are derived from the positive claims, to support multi-truth discovery.
- We present methods for re-balancing the distributions of positive/negative claims and for incorporating the influence among the co-occurring values in the same claims to improve the truth discovery accuracy. The first method neutralizes the impact of sources' behavioral features on truth discovery results and the second method improves the evaluation of values' veracity by capturing underlying correlations between values.
- We evaluate the proposed approach via experiments on various real-world datasets. The results demonstrate the effectiveness of our approach.

The rest of the paper is organized as follows. We discuss the observations that motivate our work and define the truth discovery problem in Section 2. Section 3 introduces our approach, including the probabilistic model and improvement measures. Section 4 reports our experiments and results. We discuss the related work in Section 5 and give some concluding remarks in Section 6.

2. PRELIMINARIES

Due to the correlations among the values claimed by different sources, the first step towards multi-truth discovery is to separate the evaluation of each value contained in the same claims. A characteristic of the multi-source data is that it contains only positive claims, i.e., given a specific data item, the sources only provide the values that they believe true, but do no provide the values that they believe false. While the positive claims can be used to evaluate values in terms of which values are more likely to be true, the evaluation results derived from pure positive claims represent only relative measures and a single evaluation score itself cannot indicate whether the corresponding value is true.

Incorporating mutual exclusion between distinctive values is the fundamental technique that enables truth discovery methods to acquire this ability. According to the mutual exclusion assumption, by claiming a value as the prospective truth (i.e., positive claims), a source is believed to implicitly disclaim all the other values on the same data item (i.e., negative claims). Several existing truth discovery methods [6, 16] have incorporated this assumption. Specially, all probabilistic models [20, 21] naturally conform to this assumption, as truth probability is exactly the measure that is capable of indicating the truthfulness of a value by itself.

2.1 Observations and Motivation

2.1.1 Investigation of Real-World Datasets

In view of the significance of the positive and negative claims to truth discovery, we investigate the distribution of the two types of claims in various real-world datasets, e.g., the author dataset [17], the biography dataset [11], and the movie dataset [16]. We observe imbalanced distributions of positive and negative claims in most of the investigated datasets, due to the long-tail characteristic of the real-world datasets [8, 15]:

- Most values are claimed by very few sources. For such values, the negative claims overwhelm the positive claims in number, and truth discovery methods tend to predict only partial or incomplete true values.
- A small portion of values are claimed by large numbers of sources. For such values, the positive claims overwhelm the negative claims in number, and truth discovery methods tend to predict more values than the real truth.

Intuitively, the imbalanced distribution of positive and negative claims can significantly impair the truth discovery accuracy, e.g., if a value is claimed only by a trustworthy source, it should be regarded as a true value. Unfortunately, it is very likely that a truth discovery method would recognize it as false since all the other sources (i.e., the sources that do not provide this value) are believed to disclaim this value, according to the mutual exclusion assumption. In comparison, a balanced distribution of positive/negative claims enables truth discovery methods to adapt to various problem scenarios, without being affected by the sources' behavioral features in the specific datasets. For example, if most sources in a dataset claim fewer values than the truth, the negative claims would overwhelm the positive claims in number, leading to low recall of truth discovery methods. Re-balancing the positive and negative claims could compensate the impact caused by their imbalanced distribution and therefore improve the truth discovery accuracy.

2.1.2 A Motivating Example

In the following, we illustrate the basic concepts and issues related to detecting varying numbers of true values through a running example. We will use the simplest truth discovery method, naive voting, in this example for ease of illustration; but most other methods have the similar issues.

Suppose we want to corroborate the correct authors of some books. Five data sources provide such information and only s_1 provides all true values (Table 1). The naive voting method would predict $\{Tim, Chris\}$ as the correct authors of the book (id: 0126565619), since they represent most frequently occurring claim in all sources' claims. In comparison, by separating the evaluation of distinctive values in the same claims (the results represent the positive claims, which are shown in the upper segments of Table 2), e.g., evaluating Tim and $Ellis$ separately instead of evaluating $\{Tim, Ellis\}$ together as a joint single value, naive

Table 1: An illustrative example: five sources provide information on the authors of three books. Only s_1 provides all true values.

	0132212110	0126565619	0321113586
s_1	Jeffrey; Mary; Fred	Tim; Ellis	Herb; Alex
s_2	Jeffrey	Ellis; Korper	Alex
s_3	Jeffrey	Tim; Chris	Herb; Bjarne
s_4	Mary	Tim; Ellis; Korper	Herb
s_5	Jeffrey	Tim; Chris	Herb

voting would produce $\{Tim\}$ as the truth instead. This result apparently contains only true values but the deficit is that it contains only one true value.

By incorporating mutual exclusion, the negative claims are added to the claim sets (as shown in the lower segments of Table 2). By leveraging both the positive and negative claims, the naive voting method is now able to determine the truthfulness of each value individually. Take the book (id: 0126565619) for example, since the sources that make positive claims regarding Tim are more than those making negative claims, naive voting would regard Tim as a correct author. Finally, it would predict $\{Tim, Ellis\}$ as the correct authors for the book (id: 0126565619), as the positive claims are more than the negative claims regarding each of the two values.

Although the naive method finds the correct authors of the second book, it only represents a rare case. Since most real-world sources tend to provide only partial truth, deriving the negative claims in this way could lead to imbalanced positive/negative claims, which impairs the truth discovery accuracy. For example, naive voting incorrectly recognizes $\{Mary, Fred\}$ and $Bjarne$ as false values for the first and third books, respectively. Considering source quality may helps, but cannot fully address this issue. We test different truth discovery methods on this example, but none recognizes $Fred$ as a true author of the first book since only one (s_1) of the five sources claims it to be true. Assigning higher reliability to s_1 cannot neutralize the impact of the overwhelming negative claims. On the other hand, in many cases, incorporating source quality could even deteriorate the situation. For example, many sources in the real-world only claim very few values that are frequently claim by others. On such occasions, the fewer values the source claim, the more likely it would be trusted by a truth discovery method. This mechanism further decreases the chance of the less frequently occurring values being predicted as true by truth discovery methods.

2.2 Problem Definition

We describe the key concepts and their relations in a multi-truth discovery problem (shown in Fig. 1) and explain each concept as follows.

Data Item. A data item is an object whose true values are to be detected. A data item usually exists as an attribute of a real-world entity. For example, *the author(s) of the book "Principles of Computer Architecture"* is a data item, *book* is an entity, and *author* is an attribute.

True Value. True values are the values of the data items that conform to the factual truth. Each data item may have one or more true values. For example, the true values for the authors of the book "Principles of Computer Architecture" are *Miles J. Murdocca* and *Vincent P. Heuring*.

Table 2: The example after reformatting sources' original claims into those regarding individual values: the symbol ¬ before a value means a source disclaim the value.

(a) First book

	0132212110
s_1	Jeffrey
s_1	Mary
s_1	Fred
s_2	Jeffrey
s_3	Jeffrey
s_4	Mary
s_5	Jeffrey
s_2	¬Mary
s_2	¬Fred
s_3	¬Mary
s_3	¬Fred
s_4	¬Jeffrey
s_4	¬Fred
s_5	¬Mary
s_5	¬Fred

(b) Second book

	0126565619
s_1	Tim
s_1	Ellis
s_2	Ellis
s_2	Korper
s_3	Tim
s_3	Chris
s_4	Tim
s_4	Ellis
s_4	Korper
s_5	Tim
s_5	Chris
s_1	¬Korper
s_1	¬Chris
s_2	¬Tim
s_2	¬Chris
s_3	¬Ellis
s_3	¬Korper
s_4	¬Chris
s_5	¬Ellis
s_5	¬Korper

(c) Third book

	0321113586
s_1	Herb
s_1	Alex
s_2	Alex
s_3	Herb
s_3	Bjarne
s_4	Herb
s_5	Herb
s_1	¬Bjarne
s_2	¬Herb
s_2	¬Bjarne
s_3	¬Alex
s_4	¬Alex
s_4	¬Bjarne
s_5	¬Alex
s_5	¬Bjarne

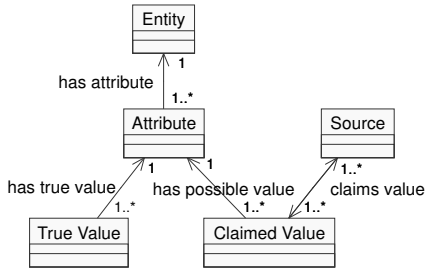


Figure 1: Relation between the key concepts in a truth discovery problem.

Data Source. It is the provider of possible true values for the data items, e.g., a website named textbookx.com, which publishes the author names of "Principles of Computer Architecture".

Claimed Value. These are the values provided by sources about the possible true values of a specific data item. Each data source may claim fewer or more values than the truth. For example, textbookx.com publishes *Miles J. Murdocca* as the only author of "Principles of Computer Architecture". In contrast, textbooksNow.com publishes two names, *Miles J. Murdocca* and *Vincent P. Heuring*, while Paperbackshop-US has no author information on this book.

Let s, o, and v be a data source, data item, and a value on this data item, respectively. We use *positive claim*, i.e., a triple (s, o, v), to represent the opinion of s that v is a true value of data item o. Similarly, we define *negative claim*, i.e., $(s, o, \neg v)$, to represent its opinion that v is a false value of o. The multi-source data are actually a list of different sources' positive claims on a set of data items. The negative claims are generated afterward for each data item and added to the claim set to facilitate multi-truth discovery.

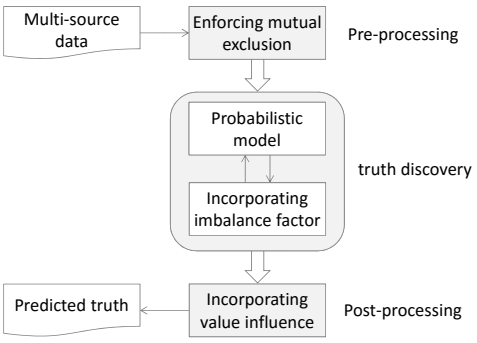


Figure 2: Overview of our approach.

The objective of the multi-truth discovery problem (MTD) is to identify the true values for each data item based on the sources' claims. The MTD distinguishes from the traditional truth discovery problem in allowing for a varying number (which is unknown *a priori*) of true values on each data item.

3. THE APPROACH

Intuitively, there are two approaches for detecting an unknown number of true values for each data item. The first approach is to evaluate the possible values and detect the number of true values (say k) separately for each data item and then predict the k values with the best evaluation results as the truth. The second approach is to employ probabilistic truth discovery models. In this paper, we address the multi-truth discovery problem based on a probabilistic approach because it naturally produces probabilistic results and has the potential of detecting the number of truth values automatically during the truth discovery process.

Our approach is based on two observations. First, separating the evaluation of individual values is just one preliminary step towards detecting multiple true values. While it on the one hand improves the truth discovery accuracy, on the other hand, it neglects the important implications in sources' claims. Second, existing truth discovery approaches do not consider the behavioral features of sources, e.g., some sources tend to claim more values than the truth while some others claim fewer values. We believe the traditional mutual exclusion definition is too strict, i.e., by claiming some value to be true, a source is believed to implicitly disclaim all the unclaimed values on the same data item.

Based on the above discussions, besides the implicit negative claims, we focus on two implications of sources' claims that are lost during the above separation, i.e., the co-occurrence of values in sources' claims, and the imbalanced distributions of positive/negative claims of sources on the data items. To emphasize on our approach's ability to deal with the possible multiple true values by combining the two aspects, we will hereafter call our approach Hybrid Multi-Truth Discovery or *MTD-hrd* for short. Fig. 2 shows the main components of *MTD-hrd*, where rectangles represent the components and the other non-shaded blocks represent the input/output information. The following subsections will introduce these components, respectively.

3.1 Enforcing Mutual Exclusion

Existing methods commonly use relative measures to evaluate values [4, 5, 6]. Given a specific data item, they first calculate a confidence score for each distinctive value and then predict the value with the highest confidence score as the truth. For the case of multiple true values, they simply take each claim as a single value and neglect the inherent correlations among the claims, which could degrade the truth discovery accuracy.

As aforementioned (Section 2), the first step of enabling the discovery of varying numbers of true values is to separate the evaluation for distinctive values for each data item. The separation allows the algorithms to capture the truth probability of values in a finer granularity and makes it possible to define a probabilistic approach that assumes conditional independence of the truth probabilities of different values for truth discovery. An example of the separation and the enforcement of mutual exclusion is shown through the transformation in Table 1 to Table 2.

3.2 Truth Discovery Model

Our probabilistic truth discovery model consists of three major elements: *value veracity*, *source reliability* and *observation* of source's claims. Fig. 3 shows the graphical structure of conditional dependence of our model, where each node in the graph represents a random variable or prior parameter. The shaded nodes indicate the observed variables and the other nodes represent latent variables. A plate with a set as its label means that the nodes within are replicated for each element in the set. A directed edge from node a to b models the conditional dependence between the two nodes in the sense that the random variable associated with b follows a probabilistic conditional distribution that takes values depending on a as parameters. Given the observed data and prior and conditional distributions, maximum *a posteriori* (MAP) estimation can be performed to find the most likely values of the unobserved variables, thus achieving truth discovery.

3.2.1 The Generative Process

In this section, we describe the model details for the three elements, respectively.

Value Veracity. We model the veracity of a value as the probability that the value is true. For each value $v \in V$, its value veracity t_v is generated from a Bernoulli distribution with parameter θ,

$$t_v \sim Bernoulli(\theta) \tag{1}$$

where t_v is a Boolean variable, and θ is the prior probability that t_v is true.

Here the prior truth probability θ determines the prior distribution of how likely each value is true. It is generated from a Beta distribution with hyperparameter $\beta = (\beta_1, \beta_2)$, where β_1 is the prior true count, and β_2 is the prior false count of each data item.

$$\theta \sim Beta(\beta_1, \beta_2) \tag{2}$$

Source Reliability. We model source reliability as the probability that a source makes correct positive and negative claims. Given a specific value, a positive (resp., negative) claim is correct when the value is actually true (resp., false). For each source $s \in S$, its reliability ϕ_s is generated from a Beta distribution with hyperparameter by $\alpha = (\alpha_1, \alpha_2)$, where α_1 is the prior false positive count, and α_2 is the prior

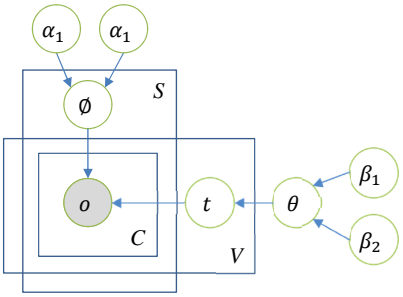


Figure 3: Graphical illustration of our probabilistic approach.

true negative count of each source:

$$\phi_s \sim Beta(\alpha_1, \alpha_2) \tag{3}$$

Observation of Sources' Claims. For each claim c of a data item o_c, suppose s_c is the source that makes this claim. The observation of c, namely X_c, is generated from a Bernoulli distribution with parameter ϕ_{s_c}

$$X_c \sim Bernoulli(\phi_{s_c}) \tag{4}$$

3.2.2 Inference Approach

Given observations on sources' claims, we perform inference to estimate the truth probabilities of values and the reliability of sources based on the above generative processes. In particular, the complete likelihood of all observations, latent variables, and unknown parameters given the hyper-parameters α and β is:

$$p(X, s, t, \theta, \phi | \alpha, \beta) = \prod_{s \in S} p(\phi_s|\alpha) \times \prod_{v \in V} \left(p(\theta_v|\beta) \theta_v^{t_v} (1-\theta_v)^{1-t_v} \mathcal{Q} \right) \tag{5}$$

where $\mathcal{Q}$ denotes the impact of sources' claims, which is the multiplication of a series of conditional probabilities, i.e.,

$$\mathcal{Q} = \prod_{c \in C} p(X_c|\phi_{s_c}) \tag{6}$$

The objective of the inference approach is to assign the appropriate truth labels to the values, so as to maximize the joint probability $p(X, s, t)$. In particular, we pursue the maximum *a posterior* (MAP) estimate of t:

$$\hat{t}_{MAP} = \arg\max_t \int p(X, s, t, \theta, \phi) \, \mathrm{d}\theta \, \mathrm{d}\phi \tag{7}$$

The joint distributions can be estimated by using various existing inference algorithms, such as the variational [1] or sampling algorithms [7]. The time complexity of these algorithms is usually $O(|C|)$ or $O(|S||V|)$, which is linear in the number of claims.

3.2.3 Calculating Imbalance Factor

The incorporation of mutual execution easily leads to imbalanced numbers of positive and negative claims over the data items and their values. The imbalance could, in turn, make truth discovery methods dependent on the sources' behavioral features in the specific datasets and thus impair the truth discovery accuracy. Basically, the probabilistic approach infers truth probabilities by maximizing the likelihood of observations of sources' claims. However, after incorporating mutual exclusion, negative claims are added and the original observations (i.e., the set of positive claims) are therefore altered. This means part of the observations used by the probabilistic approach are artificially generated and may not reflect the reality. For example, given a specific book, a cautious source may post only the names of people who are convinced to be the authors of this book, while an audacious source may post the names of all people who are possibly the authors of the book. In the first case, the source does not necessarily exclude the existence of more authors for the book; in the latter case, the source may not totally support all the authors that it claims to be true.

Given a real-world dataset, it is usually dominated by either the cautious sources or the audacious sources. The inclusion of negative claims makes truth discovery methods to, in the first case, overly decrease the truth probability of the less frequently occurring values, and in the second case, overly amplify the truth probability of all values. The first case generally leads to a low recall, while the second case usually results in low precision. Both cases lead to degraded truth discovery accuracy.

We improve the truth discovery accuracy by reducing the bias resulted from the imbalanced positive/negative distribution in the specific datasets. In particular, we distinguish between the impact of positive and negative claims and thereby rewrite Eq. (6) as:

$$\mathcal{Q} = \mathcal{Q}^+ \cdot \mathcal{Q}^- \tag{8}$$

where the positive and negative factors are defined by:

$$\begin{cases} \mathcal{Q}^+ = \prod_{c \in C^+} p(X_c|\phi_{s_c}) \\ \mathcal{Q}^- = \prod_{c' \in C^-} p(X_{c'}|\phi_{s_{c'}}) \end{cases} \tag{9}$$

To neutralize the influence of the imbalanced numbers of positive/negative claims on evaluating their respective impact, we calibrate the two types of impact by:

$$\begin{cases} \mathcal{Q}^+ = \prod_i \prod_j \sqrt[|C^+_{i,j}|]{\prod_{c \in C^+_{i,j}} p(X_c|\phi_{s_c})} \\ \mathcal{Q}^- = \prod_i \prod_j \sqrt[|C^-_{i,j}|]{\prod_{c' \in C^-_{i,j}} p(X_{c'}|\phi_{s_{c'}})} \end{cases} \tag{10}$$

where $C^+_{i,j}$ (resp., $C^-_{i,j}$) is the set of all positive (resp., negative) claims with respect to the j-th value for data item o_i.

This calibration method has two characteristics, both of which are consistent with our intuition:

- The impact of positive and negative claims are more determined by the reliability of the sources that make these claims rather than the numbers of sources that make the same claims.
- Sources of similar reliability tend to have a greater impact on determining the truthfulness of a value than the sources that have significantly different reliability.

3.3 Incorporating Value Influence

Since the values occurring in the same claims are believed to impact each other in terms of veracity, the co-occurrence of values in the same claims indicates potentially similar truth probability of these values. Based on this insight, our interpretation of value co-occurrence lies between the two extreme cases of previous truth discovery efforts: *i*) simply regarding the values in the same claim as an integral joint value—as what traditional truth discovery methods do based

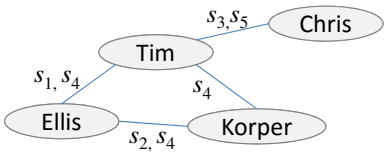


Figure 4: Association among values for the book (id: 0126565619).

on the unseparated claims, and *ii*) considering the values in the same claim as totally independent—as what the methods do based on the positive and negative claims regarding each single value.

We define weighted association among the distinctive values on the same data item to represent their influence to each other, based on which to amend the predicted truth probability of each value and to achieve better results. In particular, we represent the bipartite mapping between sources and values on each data item into a value graph, which is a graph where values are the vertices and sources are the weights of edges among the values. For example, the influence of the distinctive values of the book (id: 0126565619) in Table 1 can be represented by Fig. 4.

Note that we do not distinguish between the positive and negative influence in this graph as both types of influence are considered at the same time in our calculation. Given a data item, the influence of the distinctive values on one another can be represented as a square adjacent matrix, which should be stochastic, irreducible, and aperiodic to be guaranteed to converge to a stationary state [13]. Given a value graph, the more sources claim the co-occurrence of two values, the more relevant the values are in their truthfulness. Since the graph is already bi-directional, for the three characteristics, we initialize the weights over the edges of the graph for each data as the sum of the reliability of all sources that claim the co-occurrence of this pair of values, and then normalize the weights to ensure that every column sums to 1. We add small weights to the edges between the unconnected values to ensure a full connectivity. We finally perform iterations using page-rank methods to achieve a stationary state over the mutual influence of these values for each data item and incorporate the influence of these co-occurring values on one another in the probabilistic model by leveraging the stationary weights.

Given the veracity of values and the stationary weight between values, we calculate the veracity of an arbitrary value v by:

$$\hat{\theta}(v) = \frac{1}{1+\sum_{i=1}^{m(v)} w(v,v_i)}\theta(v) + \sum_{i=1}^{m(v)} \frac{w(v,v_i)}{1+\sum_{i=1}^{m(v)} w(v,v_i)}\theta(v_i) \tag{11}$$

where $\theta(v)$ and $\hat{\theta(v)}$ are the old and new veracity scores of v, $m(v)$ is the number of all the other values (e.g., v_i) that have an influence on v, and $w(v,v_i)$ is the stationary weight between v and v_i. Since $\sum_{i=1}^{m(v)} w(v,v_i) = 1$, Eq. (11) can be simplified as:

$$\hat{\theta}(v) = \frac{1}{2}\Big(\theta(v) + \sum_{i=1}^{m(v)} w(v,v_i)\theta(v_i)\Big) \tag{12}$$

4. EXPERIMENTS

In this section, we report the experimental studies on the comparison of our approach with the state-of-the-art algorithms using real-world datasets, and the impact of the two concerns, i.e., the co-occurrence of values in the same claim and the imbalance factors.

4.1 Experimental Setup

4.1.1 The Datasets

We used three real-world datasets in our experiments. The *author dataset* [17] contains 33,971 book-author records crawled from www.abebooks.com. Each record represents a bookstore's claim on the author(s) of a book. We removed the invalid and duplicated records, and excluded the records with only minor conflicts to make the problem more challenging. Finally, we obtained 12,473 distinctive claims describing 634 sources (i.e., websites) that provide author names on 657 books. On average, each book has 2.19 authors. The ground truth provided for the original dataset is used as gold standard.

The *biography dataset* [11] contains 11,099,730 records of people's birth and death dates, parents/children, and spouses on Wikipedia. We extracted the records related to the parent-child relation and got 2,402 people's children information edited by 54,764 users. Similar to the handling of the book-author dataset, we also removed the records with minor conflicts from this dataset. In the resulting dataset, each person has on average 2.48 children. For the experimental study purpose, we used the latest editing records as the ground truth.

We prepared the third dataset, the *director dataset*, by crawling 33,194 records from 16 major movie websites. We removed redundant records and finally obtained 6,402 movies, each on average having 1.2 directors. We sampled 200 movies and extracted their director information from citwf.com as the ground truth.

4.1.2 Evaluation Metrics

We evaluated the performance of the algorithms using three measures. For all these measures, a larger value indicates a better result.

- *Precision*, i.e., the average percentage of the predicted actual true values in the set of all predicted true values on all values of all data items.
- *Recall*, i.e., the average percentage of the predicted actual true values in the set of all actually true values on all values of all data items.
- F_1 *score*, i.e., the harmonious mean (i.e., a weighted average) of *precision* and *recall* ranging from 0 to 1. It is calculated using:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{13}$$

4.1.3 Baseline Methods

We compared our approach with three categories of truth discovery methods. All the algorithms are unsupervised and some traditional truth discovery algorithms (i.e., *Voting* and

Table 3: Comparison of different algorithms on the three datasets: the best performance values are bolded, where *precision*, *recall* and F_1 *score* are all in the range of [0,1].

Method	Author dataset			Biography dataset			Director dataset		
	Precision	Recall	F_1 score	Precision	Recall	F_1 score	Precision	Recall	F_1 score
Voting	**0.88**	0.23	0.36	**0.90**	0.12	0.21	**0.91**	0.15	0.26
Sums	0.69	0.49	0.57	0.76	0.59	0.66	0.85	0.64	0.73
2-Estimates	0.79	0.65	0.71	0.80	0.62	0.70	0.87	0.77	0.82
TruthFinder	0.73	0.78	0.75	0.80	0.89	0.84	0.85	0.88	0.86
Voting-N	0.79	0.49	0.60	0.84	0.59	0.69	0.87	0.64	0.74
Sums-N	0.73	0.65	0.69	0.77	0.52	0.62	0.81	0.56	0.66
Accu-N	0.73	0.78	0.75	0.77	0.82	0.79	0.85	0.77	0.81
LTM	0.84	0.78	0.81	0.84	0.82	0.83	0.81	0.83	0.82
MBM	0.79	**0.87**	0.83	0.87	0.85	0.86	0.87	0.86	0.86
Our approach	**0.88**	**0.87**	**0.87**	**0.90**	**0.87**	**0.88**	0.87	**0.89**	**0.88**

Sums) were improved in different ways to cope with the multi-truth scenario.

Traditional Truth Discovery Methods. The methods of this category consider all the values of the each claim as a single value. We selected several typical and competitive algorithms from this category for the experimental comparison. In the following, both *Voting* and *Sums* were modified by taking into account of the mutual exclusion consideration.

- *Voting*: for each item, the value claimed by the most sources is produced as the estimated truth. This method regards a value as true if the proportion of the sources that claim the value exceeds a certain threshold.
- *Sums*: this method evaluates sources and values alternately from each other by iteration. It computes the total reliability of all sources that claim and disclaim a value separately and recognizes the value as true if the former is larger than the latter.
- *2-Estimates* [6]: this method assumes that claiming one value indicates disclaiming (vote against) all unclaimed values. A value is recognized as true if its veracity score exceeds 0.
- *TruthFinder* [17]: this method considers inter-value influence and evaluates each value by a probability. TruthFinder recognizes a value as true if its veracity score exceeds 0.5.

Existing Multi-Truth Discovery Methods. There are two multi-truth methods, discussed as the following:

- LTM [20]: this method assumes prior distributions of latent variables and thereby constructs a probabilistic graphical model to infer sources' reliability and values' veracity.
- MBM [16]: this method deals with the multi-truth problem by incorporating new mutual exclusion definition.

Improved Single-Truth Discovery Methods. Both *Voting* and *Sums* can be improved using a different approach, i.e., incorporating them with true value number prediction methods, to cope with the multi-truth problem. The original *Accu* method [4] evaluates a value by the *a posterior* probability using the Bayesian model without the variable distribution assumptions. As a result, *Accu* can be improved in a similar way to cope with multi-truth discovery.

In particular, we used the truth discovery procedures of the original methods to produce evaluations to the values and followed a similar procedure to predict the number of true values. Suppose the estimated true value number is N, the values with the top-N highest evaluation values are predicted as the true values. We denote the three methods improved by the above approaches by *Voting-N*, *Sums-N*, and *Accu-N*, respectively.

Our approach. Our approach combines the consideration of two aspects of concern, the co-occurrence of values and imbalance in sources' claims, in the same probabilistic model.

We excluded comparison with the methods in [11], which are inapplicable for discovering varying numbers of true values. In particular, the approach requires normalizing the veracity of values, which is infeasible for our problem. Besides, the methods in [19] focus on handling numerical data, while our approach is proposed specially for categorical data.

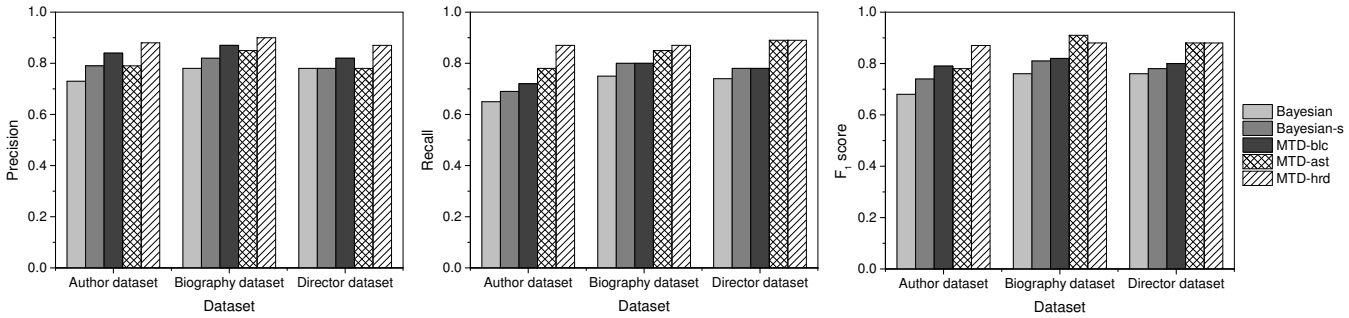
4.1.4 Algorithm Configuration

We implemented all algorithms using Java SDK 7 and conducted experiments on a 64-bit Windows 7 PC with an octa-core 3.4GHz CPU and 8GB RAM.

To ensure fair comparisons, we first ran a series of experiments to decide the optimal parameter settings for the baseline methods. For our method, we initialized the prior accuracy of all sources to 0.9. We set the parameters in a manner that makes sure the actual prior counts should be at the same scale as the number of facts to ensure its effectiveness. Since all the three datasets are at the scale of ten thousands, we defined the same parameters for the datasets, i.e., $(\alpha_1, \alpha_2) = (50, 50)$ and $(\beta_1, \beta_2) = (10, 10)$. For the both sets of parameters, we defined relatively small uniform prior, which enforced no prior bias towards the truth discovery results. Note that we used 0.5 other than other values as the threshold to determine the truthfulness of values in many of the compared algorithms, since tuning this value requires using ground truth and makes the comparison unfair to the other algorithms.

4.2 Comparative Studies

Table 3 shows the performance of different algorithms on the three datasets in terms of precision, recall, and the F_1 score. From the table, we can see that our approach consistently achieves the best (or the second best) precision among all the compared methods. The majority voting method always achieves the best precision. This indicates that some true values are claimed by most data sources in all the three datasets. All algorithms achieve lower precision on the *au-*



(a) Precision **(b)** Recall **(c)** F_1 score

Figure 5: Comparison of the five algorithms in terms of precision, recall, and F_1 score on three real-world datasets.

thor dataset because we intentionally eliminated the records with minor conflicts when preparing the dataset to make the problem more challenging.

For recall, TruthFinder, MBM, and our approach generally achieve better recall over all the other baselines. Among the inferior baseline methods, those based on the consideration of mutual exclusion neglect the uncertainty in inferring the sources' negative claims from the raw data, i.e., the positive claims, while the others based on the prediction of true value numbers suffer from the uncertainty in the predicted true value numbers. Another problem with the methods based on the prediction of true value numbers is that sources' reliability is separately evaluated in predicting the true value numbers and in evaluating the different values. Consequently, the sources' reliability used to predict true value numbers may not be consistent with the ranked evaluation results of the values, thus degrading the recall.

It should be noted that TruthFinder achieves better recall yet generally lower precision than the other baselines, which may attribute to its overestimation of veracity scores. Majority voting achieves the lowest recall on all the three datasets. This is because most sources tend to provide only a minor proportion of the entire truth. Given an arbitrary value, the sources that do not claim this value almost always overwhelm the sources that claim this value. As a result, after incorporating the mutual exclusion consideration, majority voting could hardly yield a bigger proportion value than 0.5 due to the imbalanced distribution of positive and negative claims. This may imply that the majority voting method should be better used in a supervised or semi-supervised manner where the threshold (i.e., 0.5 in the experiments) can be tuned for the specific dataset based on sufficient ground truth.

The comparison results with respect to *Voting vs. Voting-N* and *Sums vs. Sums-N* show no significant difference between the traditional methods improved by the two approaches (i.e., incorporating either the mutual exclusion and incorporating the potential number of true values), except for the majority voting method, whose recall is improved significantly by predicting the true value numbers. The reason lies in that the approach based on the true value number prediction is not affected by the imbalanced distribution of positive and negative claims.

The comparison results help us better understand the different approaches for incorporating the multi-truth concerning in addressing the true discovery problem. Considering the mutual exclusion relation can help detecting any number of true values naturally. However, due to the strong assumption of implicit negative claims, it easily suffers from the imbalanced distribution of the positive and implicit negative claims. Predicting the potential number of true values can avoid the above problem. However, it produces an inconsistent estimation of source reliability, which cannot be easily addressed. Besides, both approaches require separating the concerns of different values in the same claim, which cause the loss of the implicit correlation between the co-occurring values. Our approach takes into account both aspects of concern about the imbalanced claim distribution and the implicit correlation between values. The good performance of our approach in the comparison shows the effectiveness of incorporating these concerns for addressing the truth discovery problem with unknown numbers of true values.

4.3 Impact of Different Concerns

To evaluate the impact of different concerns (i.e., the separate consideration of different values in the same claim, the imbalanced distribution of positive and implicit negative claims, and the co-occurrence of values in the same claim), we implemented two algorithms based on the Bayesian probabilistic model.

- *Bayesian*: this is the basic graphical probabilistic model that calculates truth probabilities by inferring a series of latent variables. The values contained in each single claim is defined as a variable. The values contained in the claim that has the highest truth probability are regarded as truth.
- *Bayesian-s*: this model differs from *Bayesian* in defining each single value as a variable. Those values with a truth probability over 0.5 are regarded as truth.

We further derived two variants of our approach (i.e., *MTD-hrd*) for the comparison:

- *MTD-blc*: a version of *MTD-hrd* that only adopts the imbalance factor to improve the basic probabilistic model.
- *MTD-ast*: a version of *MTD-hrd* that only adopts the influence between the different values of each data item to improve the basic probabilistic model.

Fig. 5 shows the comparison of the above five implementations of the probabilistic truth discovery approach. Results on all the three datasets show that when the actual number

of true values is unknown *a priori*, separating the concerns of individual values in the same claim can improve both the precision and recall of truth discovery, as indicated by the better performance of *Bayesian-s* than *Bayesian*. Given that different sources' claims about the same data item may overlap, the truth probabilities of the sources' claims about the same data item are neither statistically nor conditionally independent. Due to this reason, traditional truth discovery methods, even some might be applicable, are generally unsuitable for detecting unknown numbers of true values if unmodified.

The superior precision and recall of *MTD-blc* and *MTD-ast* comparing to *Bayesian-s* indicate that both measures—incorporating the imbalance factor and incorporating the value influence—can help improve the accuracy. It is worth noting that the first measure has a greater impact on the precision while the second on the recall of truth discovery. Our approach, which incorporates both measures, always achieves the best precision and recall, as well as F_1 score, when compared with the other four methods.

To investigate the impact of different concerns to the scalability of the probabilistic approach, we conducted both theoretical analysis and experimental studies to compare the five algorithms. The time complexity of truth discovery is usually determined by the number of connections between sources and truth variables, which is no larger than $M \times N$, where M and N are the numbers of sources and truth variables, respectively. For *Bayesian*, the truth variables are the claims. For all the other four algorithms, truth variables are individual values. Considering the upper-bound notations, the time complexity of *Bayesian* is $O(M \times |C|)$, where $|C|$ is the maximum number of claims for each source. In contrast, the time complexity of all the other algorithms is $O(M \times |V|)$, where $|V|$ is the maximum number of values claimed for each source. Because each source can make at most one claim on each data item, we have $|C| \leq |V|$.

Fig. 6 shows the performance of the algorithms in terms of computation time on the director dataset. The results show that evaluating the values separately may not significantly affect the efficiency but incorporating the value influence does. In particular, incorporating value influence (as illustrated in Section 3.3) might be time-consuming as it requires achieving the stationary state of graph weights using page-rank algorithms. To better evaluate this impact, we further compared *MTD-hrd* and *MTD-ast* with their respective simplified versions, which directly use the normalized graph weights instead of the stationary weights to calculate the influence. The results show only a slight difference in the resulting accuracy of using and not using the page-rank methods, while the efficiency could be drastically improved through the simplification. The comparison results on other datasets show the similar results.

5. RELATED WORK

Due to the significance of data veracity, tremendous efforts have been conducted on truth discovery [10]. The truth discovery problem is first coined by Yin *et al.* in 2008 [17]. It is defined as the process of finding the true values of a set of data items from the conflicting records reported by different sources. The problem is non-trivial because most real-world sources are unreliable, and the ground truth is difficult to obtain and generally insufficient to support a supervised approach.

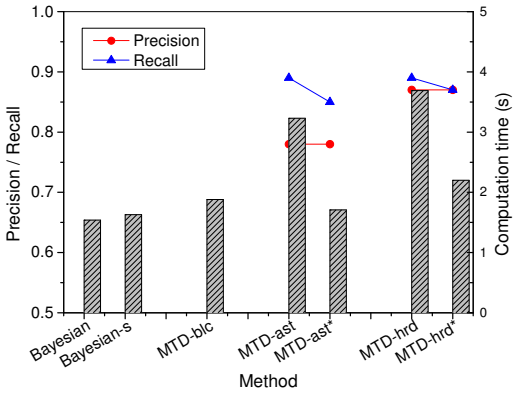


Figure 6: Performance comparison of the algorithms: all the five algorithms are compared in terms of computation time. Only *MTD-ast* and *MTD-hrd* are compared with their simplified variations, i.e., *MTD-ast and *MTD-hrd**, in terms of precision and recall.**

Source quality is an important factor that determines the trustworthiness of information. Generally, a source of high quality has a higher possibility of providing correct information than a low-quality source. Therefore, most truth discovery methods jointly evaluate the trustworthiness of information and source quality through an iterative process, where the two aspects are estimated alternately from each other. Those methods diverge into different categories according to the techniques used for the estimation, such as the iterative approach [17, 11, 6, 4], the Maximum Likelihood Estimation (MLE) approach [14], optimization approach [18, 9, 8], and the probabilistic approach [20, 19, 12].

Despite active research in the field, a fundamental problem remains unresolved, i.e., most existing research is confined by the assumption that each data item has only one true value. While these methods commonly obtain the single true value by ranking the evaluation results of all distinctive values, they cannot automatically detect the number of true values and thus cannot be applied to the multi-truth scenario. In fact, it is common in the real world that an item has more than one true value. Failure to incorporate this feature may degrade the truth discovery accuracy.

Unfortunately, until now, there are very few works on multi-truth discovery. Intuitively, traditional truth discovery methods can be used for discovering multi-truth by regarding all the values in the same claim as a single value. But the problem is that this modification approach does not consider the relationship between different claims, i.e., some claims may overlap in their contained values, and therefore lead to low accuracy. In [20], Zhao *et al.* propose a Bayesian model that is compatible with the multi-truth scenario, which shows a positive effect on the accuracy and allows for the existence of multiple true values. In [16], Wang *et al.* define a new *mutual exclusive relation* among values for multi-truth discovery, and incorporate sources' confidence on their claims and a finer-grained copy detection technique into a Bayesian framework to address the problem. Both above methods avail from using the probabilistic methods, which can automatically detect the number of true values as a byproduct. However, they consider each distinctive value independently, while on the other hand, neglect their inter-relations. We believe the co-occurrence of values

in the same claim reflects sources' belief on their similarity in the truthfulness. Zhi *et al.* [21] also improve the mutually exclusive relation, but in a way that models the silence rate of sources to allow for the possible non-existence of true values in the multi-source data. However, this work does not consider the multi-truth scenario.

Our interpretation of the computing of values lies between two extreme conditions, namely i) simply regarding the values in the same claim as an integral joint value, and ii) considering the values in the same claim totally separately. We believe that separating the consideration of different values in the same claim is only a first step, which enables existing approaches to deal with multi-truth discovery problem. A further step should include reconsidering the important implications contained in the problem inputs and developing approaches that can leverage these implications to address the problem effectively. Our approach is proposed based on the above approaches and considerations, but it is more generic in terms of allowing for the existence of a flexible number (e.g., zero, single, or multiple) of true values for each data item and being able to deal with imbalanced source claims, which form the main contributions of this paper.

6. CONCLUSION

In this paper, we focus on the problem of detecting varying numbers of true values on different data items from multi-source data, or the multi-truth discovery problem (MTD). Although the data items with multiple true values widely exist in the real world, the MTD is rarely studied by previous efforts. We have proposed a probabilistic model, which comprehensively incorporates implications from sources' claims in the multi-source data to address the MTD. In particular, we separate the evaluation of each value in the same claims and infer the implicit negative claims of sources to enable multi-truth discovery by a probabilistic approach. We further incorporate two measures, 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 truth discovery accuracy. The resulting approach is able to perform accurate multi-truth discovery under various conditions of sources' behavioral features manifested by the dataset. Experimental results on three real-world datasets demonstrate the effectiveness of our approach.

7. REFERENCES

[1] M. J. Beal. *Variational algorithms for approximate bayesian inference*. University of London, 2003.

[2] D. Benslimane, Q. Z. Sheng, M. Barhamgi, and H. Prade. The uncertain web: concepts, challenges, and current solutions. *ACM Trans. Internet Technol. (TOIT)*, 16(1):1, 2015.

[3] C. Dobre and F. Xhafa. Intelligent services for big data science. *Future Gener. Comput. Syst.*, 37:267–281, 2014.

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

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

[6] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *Proc. ACM Intl. Conf. on Web Search and Data Mining (WSDM)*, pages 131–140, 2010.

[7] A. E. Gelfand and A. F. Smith. Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, 85(410):398–409, 1990.

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

[9] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proc. ACM SIGMOD Intl. Conf. on Mgmt. of Data*, pages 1187–1198, 2014.

[10] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. A survey on truth discovery. *ACM SIGKDD Explor. Newsletter*, 17(2):1–16, 2016.

[11] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *Proc. Intl. Conf. on Comput. Linguistics (COLING)*, pages 877–885, 2010.

[12] J. Pasternack and D. Roth. Latent credibility analysis. In *Proc. Intl. World Wide Web Conf. (WWW)*, pages 1009–1020, 2013.

[13] A. Rajaraman, J. D. Ullman, J. D. Ullman, and J. D. Ullman. *Mining of massive datasets*, volume 77. Cambridge University Press, 2012.

[14] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher. On truth discovery in social sensing: a maximum likelihood estimation approach. In *Proc. ACM Intl. Conf. on Info. Processing in Sensor Networks (Sensys)*, pages 233–244, 2012.

[15] X. Wang, Q. Z. Sheng, X. S. Fang, X. Li, X. Xu, and L. Yao. Approximate truth discovery via problem scale reduction. In *Proc. ACM Intl. Conf. on Info. and Knowl. Mgmt. (CIKM)*, pages 503–512, 2015.

[16] X. Wang, Q. Z. Sheng, X. S. Fang, L. Yao, X. Xu, and X. Li. An integrated bayesian approach for effective multi-truth discovery. In *Proc. ACM Intl. Conf. on Info. and Knowl. Mgmt. (CIKM)*, pages 493–502, 2015.

[17] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE Trans. Knowl.and Data Eng. (TKDE)*, 20(6):796–808, 2008.

[18] X. Yin and W. Tan. Semi-supervised truth discovery. In *Proc. Intl. World Wide Web Conf. (WWW)*, pages 217–226, 2011.

[19] B. Zhao and J. Han. A probabilistic model for estimating real-valued truth from conflicting sources. In *Proc. Intl. Workshop on Quality in DataBases (QDB), coheld with VLDB*, 2012.

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

[21] S. Zhi, B. Zhao, W. Tong, J. Gao, D. Yu, H. Ji, and J. Han. Modeling truth existence in truth discovery. In *Proc. ACM SIGKDD Intl. Conf. on Knowl. Discov. and Data Mining (WSDM)*, pages 1543–1552, 2015.