# From Appearance to Essence: Comparing Truth Discovery Methods without Using Ground Truth

XIU SUSIE FANG, Donghua University, China
QUAN Z. SHENG, Macquarie University, Australia
XIANZHI WANG, University of Technology Sydney, Australia
WEI EMMA ZHANG, University of Adelaide, Australia
ANNE H. H. NGU, Texas State University, United States
JIAN YANG, Macquarie University, Australia



Truth discovery has been widely studied in recent years as a fundamental means for resolving the conflicts in multi-source data. Although many truth discovery methods have been proposed based on different considerations and intuitions, investigations show that no single method consistently outperforms the others. To select the right truth discovery method for a specific application scenario, it becomes essential to evaluate and compare the performance of different methods. A drawback of current research efforts is that they commonly assume the availability of certain ground truth for the evaluation of methods. However, the ground truth may be very limited or even impossible to obtain, rendering the evaluation biased. In this article, we present *CompTruthHyp*, a generic approach for comparing the performance of truth discovery methods without using ground truth. In particular, our approach calculates the probability of observations in a dataset based on the output of different methods. The probability is then ranked to reflect the performance of these methods. We review and compare 12 representative truth discovery methods and consider both single-valued and multi-valued objects. The empirical studies on both real-world and synthetic datasets demonstrate the effectiveness of our approach for comparing truth discovery methods.

CCS Concepts: • **Information systems** → **Data mining**; **Retrieval tasks and goals**;

Additional Key Words and Phrases: Web search, truth discovery methods, sparse ground truth, performance evaluation, single-valued objects, multi-valued objects

**ACM Reference format:**
Xiu Susie Fang, Quan Z. Sheng, Xianzhi Wang, Wei Emma Zhang, Anne H. H. Ngu, and Jian Yang. 2020. From Appearance to Essence: Comparing Truth Discovery Methods without Using Ground Truth. *ACM Trans. Intell. Syst. Technol.* 11, 6, Article 74 (September 2020), 24 pages.
https://doi.org/10.1145/3411749

Q. Z. Sheng's work has been partially supported by Australian Research Council (ARC) Future Fellowship Grant FT140101247 and Discovery Project DP200102298.

Authors' addresses: X. S. Fang, Donghua University, China; email: xiu.fang@dhu.edu.cn; Q. Z. Sheng, Macquarie University, Australia; email: michael.sheng@mq.edu.au; X. Wang, University of Technology Sydney, Australia; email: xianzhi.wang@uts.edu.au; W. E. Zhang, University of Adelaide, Australia; email: wei.e.zhang@adelaide.edu.au; A. H. H. Ngu, Texas State University, United States; email: angu@txstate.edu; J. Yang, Macquarie University, Australia; email: jian.yang@mq.edu.au.


2157-6904/2020/09-ART74 $15.00
https://doi.org/10.1145/3411749

## 1 INTRODUCTION

The World Wide Web has become a platform of paramount importance for storing, collecting, processing, querying, and managing the Big Data in recent years, with around 2.5 quintillion bytes of data created every day through various channels such as blogs, social networks, discussion forums, and crowd-sourcing platforms.[1] People from various domains, such as medical care, government, and business, are relying on these data to fulfill their information needs. Information about the same objects can often be collected from a variety of sources. However, due to the autonomous nature of Web sources, conflicts may be reported among different Web sources. To help users determine the veracity of multi-source data, a fundamental research topic, *truth discovery*, has attracted broad attentions recently [1, 38].

So far, various truth discovery methods [14, 24, 44] have been proposed based on different considerations and intuitions. However, investigations show that no methods could constantly outperform the others in all application scenarios [22, 24, 37]. Moreover, Li et al. [22] demonstrate with experiments that even an improved method does not always outperform its original version, such as *Investment* and *PooledInvestment* [31], *Cosine, 2-Estimates,* and *3-Estimates* [16]. Therefore, to help users select the most suitable method to fulfill their application needs, it becomes essential to evaluate and compare the performance of different methods.

To evaluate the effectiveness of truth discovery methods, current research usually measures their performance in terms of *accuracy* (or *error rate*), *F1-score*, *recall*, *precision*, *specificity* for categorical data [37], and *Mean of Absolute Error* (MAE) and *Root of Mean Square Error* (RMSE) for continuous data [24]. All these metrics are measured and compared based on the assumption that a reasonable amount of ground truth is available. However, the fact is, the labor cost of ground truth collection is rather expensive. Ground truth is often very limited or even impossible to obtain (generally less than 10% of the size of the original dataset [37]). For example, the knowledge graph construction [8] involves a large number of objects, making it impossible to have even a small set of ground truth, which requires enormous human efforts.

The lack of sufficient ground truth can, in many cases, statistically undermine the legitimacy of evaluating and comparing existing methods using the ground truth–based approach. Previous comparative studies [5, 6, 21, 22, 25, 34, 40, 41, 44, 49, 50] are all based on real-world datasets with sparse ground truth, which could bring biases to the performance evaluation of methods. Methods with good accuracy may, by chance, return incorrect results on the particular objects covered by the sparse ground truth, while methods with poor accuracy may, occasionally, be consistent with the sparse ground truth. Moreover, methods that show the same accuracy on the rather limited objects covered by the sparse ground truth may have different performance in reality.

Under this circumstance, it is hard to conclude which method performs better or which method performs best for specific application scenarios, as the comparison results cannot be fully trusted. Therefore, evaluating the performance of various truth discovery methods with missing or very limited ground truth can be a significant and challenging problem for the truth discovery applications [24]. We identify the key challenges around this issue as the following:

- The only way to obtain evidence for performance evaluation without ground truth is to extract features from the given dataset for truth discovery [22, 24, 37]. However, the features of a dataset are sometimes complex, encompassing source-to-source, source-to-object, object-to-value, and value-to-value relations. In addition, it is challenging to find a method to capture those relations without creating additional biases.

[1] https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/.

- Current truth discovery methods commonly determine value veracity and calculate source trustworthiness jointly. Source trustworthiness and value confidence scores are the common intermediates of the existing methods, which are also the key elements for identifying the truth for each object [24]. Therefore, we can consider identifying the relations among sources, objects, and values by leveraging those measurements to match the relations extracted from the given dataset. However, even if we are able to obtain the features of the given dataset, different truth discovery methods may calculate the source trustworthiness and value confidence scores using different metrics, which have various meanings and require non-trivial normalization.
- Even if we are able to resolve the above two challenging issues, it is still tricky to find appropriate metrics for comparing those features to fulfill the requirement of method comparison.

In this article, we focus on truth discovery method comparison without using ground truth. In a nutshell, we make the following main contributions:

- To our knowledge, we are the first to reveal the bias introduced by sparse ground truth in evaluating the truth discovery methods by conducting experiments on synthetic datasets with different coverages of the leveraged ground truth.
- We analyze, implement, and compare 12 specific truth discovery methods, including *majority voting, Sums, Average-Log, Investment, PooledInvestment* [31], *TruthFinder* [45], *2-Estimates, 3-Estimates* [16], *Accu* [6], *CRH* [21, 25], *SimpleLCA*, and *GuessLCA* [33].
- We propose a novel approach, called *CompTruthHyp*, to **comp**are the performance of **truth** discovery methods without using ground truth by considering the output of each method as a **hyp**othesis about the ground truth. CompTruthHyp takes both single-valued and multi-valued objects into consideration. It utilizes the output of all methods to quantify the probability of observation of the dataset and then determines the method with the largest probability to be the most accurate.
- We conduct extensive experiments on both synthetic and real-world datasets to demonstrate the effectiveness of our proposed approach. Our approach consistently achieves more accurate rankings of the 12 methods than traditional ground truth–based evaluation approach.

The rest of the article is organized as follows: We review the related work in Section 2. Section 3 introduces some background knowledge about truth discovery and the observations that motivate our work. Section 4 presents our approach. We report our experiments and results in Section 5. Section 6 provides some concluding remarks.

## 2 RELATED WORK

Due to the significance of the veracity of the *Big Data*, truth discovery has been a hot topic and studied actively over past few years in the database community [11, 12]. The *primitive methods* are typically *rule-based*, i.e., *voting* and *averaging*. For categorical data, people predict the values with the highest number of occurrences as the truth, while for the continuous data, they naively take the *mean* as the true values. These methods make the assumption that sources are equally reliable, thus they show low accuracy for the cases that many sources provide low-quality data. To relax the assumption voting or averaging makes, Yin et al. [45] first formulated the truth discovery problem in 2008. In their work, a Bayesian-based heuristic algorithm is proposed, which computes the probability of each claim being correct given the estimated source weights and the influences between claims. After that, 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 (see References [2, 22, 24, 37, 51] for relevant surveys).

We can roughly classify the methods into five groups. The *iterative* methods [16, 31, 45] iteratively calculate value veracity and source reliability from each other until certain convergence condition is met. The *Bayesian point estimation* methods [6, 40] adopt *Bayesian analysis* to compute the maximum *a posteriori* probability or *MAP* value for each object. The *link*-based methods [19, 31] 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. The *probabilistic graphical model*–based methods [41, 49, 50] apply probabilistic graphical models to jointly reason about source trustworthiness and value correctness. Finally, the *optimization*-based methods [20, 21] formulate the truth discovery problem as an optimization problem.

Different methods have different assumptions about input data, source reliability, relations among sources and objects, claimed values, identified truths, and take different unique characteristics of different application domains into consideration. For input data, the works in References [10, 27, 46] assume that a small set of truths is available and thus the proposed algorithms work in a semi-supervised setting. For source reliability and relations among sources, most methods [6, 16, 21, 31, 45, 46, 49] make the source consistency assumption that a source is likely to provide true information with the same probability for all the objects. Some methods [16, 21, 31, 45, 49] make the source independence assumption that data sources are independent of each other, i.e., no source-to-source relationship exists in the given dataset. To relax the source-independence assumption, the observations of sources' authority features and sources' copying relations have been presented in Reference [19] and References [4–7, 23, 34, 48]. 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 works adopt an external trustful source [8], a subset of labeled data [10, 27, 46], the similarity among sources [47], or the two-sided source graph [13, 14] as prior knowledge to initialize or help initializing the source reliability. A neural network approach that learns complex relational dependency between source reliability and claim truthfulness has been recently proposed for truth discovery in social sensing [28]. In some scenarios, it is reasonable to estimate multiple source reliabilities for a single source so the variety in source reliability can be captured. Therefore, several methods [18, 26, 46] have been designed to capture the fine-grained source reliability. In most truth discovery works, source reliability is a parameter that is positively correlated with the probability of a source asserting truths. However, in References [20, 30, 33, 36, 50], the meaning of this parameter is further enriched to fit more complex application scenarios. The research works in References [14, 20, 43] take the long-tail phenomenon on source coverage into consideration to avoid small sources from being assigned extreme reliability. By considering incorporation of bad sources may even hurt the performance of truth discovery, References [4, 10, 36] provide methods to wisely select sources for truth discovery constrained by the cost and output quality. For relations among objects, in References [16, 31, 47], how object difficulty and the relations among objects affect truth discovery have been taken into consideration. A series of methods incorporate the implications or issues of the claimed values in the development of truth discovery, such as data type (e.g., GTM [49], CATD [20], EvolvT [52] are designed for continuous data, while 2-Estimates and 3-Estimates [16], Investment, and PooledInvestment [31] are designed for categorical data), missing values [39], complementary vote [16, 50], and Local Closed World Assumption (LCWA) [8, 9], value similarity [6, 45], hierarchical structure of claimed values [8, 9]. For identified truths, as multi-valued objects widely exist in the real world, various works have also been proposed to resolve the challenges presented by the multi-truth discovery (MTD) [14,

15, 34, 38, 40–42, 50, 53]. Instead of providing a point estimator for each object's truth, Xiao et al. [44] propose a novel truth discovery method (i.e., ETCIBoot) to construct confidence interval estimates as well as identify truths, where the bootstrapping techniques are nicely integrated into the truth discovery procedure.

Generally, there are two categories of previous studies on performance evaluation and comparison of truth discovery methods. The first category includes the works that propose novel and advanced approaches in various scenarios. To validate the performance of their proposed approaches and show how their approaches outperform the state-of-the-art methods, those projects conduct comparative studies by running experiments on real-world datasets with manually collected ground truth. Yin et al. [45] show the effectiveness of their proposed *TruthFinder* by conducting experiments on one real-world dataset, i.e., *Book-Author* dataset, which contains 1,263 objects. The manually collected ground truth only covers 7.91% of the objects. With truth discovery gaining growing popularity, considerable methods [5, 10, 14, 15, 20, 25, 29, 31, 35, 40, 41, 43, 44, 49, 50, 51] have been proposed to deal with various scenarios. Those works, however, have the common limitation that they either require labor-intensive labelling of data or use datasets with limited ground truth to conduct experiments. Besides the Book-Author dataset, the frequently-used datasets, including *Flight* [22] (covering 8.33% of complete ground truth), *Population* [31] (0.702%), *Movie* [50] (0.663%) and *Biography* [31] (0.069%) are all feature sparse or have low-quality ground truth, which makes the experimental data evaluated on those datasets cannot be fully trusted. The game dataset [20, 44] is collected by crowd-sourcing, which contains the answers of 2,103 questions from 37,029 Android users based on a TV game show "Who Wants to Be a Millionaire" via an Android App. This type of datasets are usually limited to specific sets of questions and require a high labor cost.

The second category of the studies is presented in References [22, 24, 37, 51], which aim at investigating and analyzing the strengths and limitations of the current state-of-the-art techniques. In particular, Li et al. [22] study the performance of 16 data fusion methods, in terms of precision and recall, on two real-world domains, namely, *Stock* and *Flight*. Based on their experiments, the authors point out that the collected ground truth tends to trust data from certain sources, which sometimes puts wrong values or coarse-grained values in the ground truth. Moreover, we find that their constructed ground truth is relatively sparse, with the one for the stock domain covering only 200/1,000 = 20% of the complete ground truth, and the one for the flight domain covering only 100/1,200 = 8.33%. The most recent survey [24] provides a comprehensive overview of truth discovery methods and summarizes them from five different aspects, but they do not conduct any comparative experiments to show the diverse performance of the methods. Waguih et al. [37] point out that the sparse ground truth is not statistically significant to be legitimately leveraged for the accuracy evaluation and comparison of methods. To the best of our knowledge, they are the first to implement a dataset generator to generate synthetic datasets with the control over ground truth distribution for the sake of comparing existing methods. Different from their work, our approach tries to evaluate the performance of various truth discovery methods without using ground truth, which is applicable to more general real-world scenarios.

## 3 PRELIMINARIES

### 3.1 Problem Formulation

Current truth discovery methods take as input some conflicting triples (i.e., a given dataset) in the form of {*source, object, value*}, where *source* ($s \in S$) denotes the location where the data originates, *object* ($o \in O$) is an attribute of an entity, and *value* ($V_{s_o} \subset V$) depicts the potential value set of an object claimed by a source. For example, a triple, {"www.imdb.com," "the director of Beauty and the Beast," "Bill Condon"}, indicates that the website "IMDb" claims that the director of the movie

Table 1. Notations Used in the Article

| Notation | Explanation |
|---|---|
| $o$, $O$ | An object (respectively, Set of all objects), $o$ may be single-valued/multi-valued |
| $s$, $S$ | A source (respectively, Set of all sources) |
| $v$, $V$ | A claimed value (respectively, a set of all claimed values) |
| $V_s$ | The set of all values provided by $s$ |
| $V_o$ | The set of all claimed values on $o$ |
| $m$, $M$ | A truth discovery method (respectively, Set of truth discovery methods) |
| $V_{s_o}$ | The potential value set of $o$ claimed by $s$ |
| $V_o{}^*$, $V^*$ | The ground truth of $o$ (respectively, of the given dataset) |
| $V_o{}^m$, $V^m$ | The identified truth of $o$ (respectively, the given dataset) output by $m$ |
| $V^i$ | The incomplete ground truth of the given dataset |
| $S_v$ | The set of sources provide claimed value $v$ on an object |
| $c_{\mathcal{V}}$ | The confidence score of $\mathcal{V}$, $\mathcal{V}$ is a single joint value |
| $\tau_s$ | The trustworthiness of $s$ |
| $\phi$ | The observation of which value each source in the given dataset votes for |
| $\phi_{s_v}$ | The observation of $s$ providing a particular value $v$ ($v \in V_o$) |
| $\phi_s$ | The observation of source $s$ with its claimed values |
| $P(\phi \mid V^m)$ | The probability of $\phi$ conditioned on $V^m$ |
| $\tau_s(m)$ | Given $V^m$, the probability that the claimed values of $s$ is true |
| $P_s(v_t \mid V_o^m)$ (respectively, $P_s(v_f \mid V_o^m)$) | Given $V_o^m$, the probability $s$ provides a particular true (respectively, false) value on $o$ |
| ${V_s}^t(m)$, ${V_s}^f(m)$ | The set of all true (respectively, false) values provided by $s$, given $V^m$ |
| $P(\phi_{s_v} \mid V^m)$ | The probability of $\phi_{s_v}$ conditioned on $V^m$ |
| $P(\phi_s \mid V^m)$ | The probability of $\phi_s$ conditioned on $V^m$ |
| $C_m$ | The *confidence* of method $m$ |

"Beauty and the Beast" is "Bill Condon." If $o$ is a single-valued object, then $|V_{s_o}| = 1$. For example, "the age of a person" only has one single value; however, if $o$ is a multi-valued object, $|V_{s_o}|$ is bigger than 1. For example, a person might have more than one child.

Truth discovery methods infer truth labels ("true" or "false") for the triples as the output. According to whether the methods assume more than one true value for each object [50], the current methods can be grouped into two categories: single-valued methods [6, 16, 21, 31–33, 45] and multi-valued methods [14, 15, 34, 38, 40–42, 50, 53]. Single-valued methods infer a truth label to each triple. When multi-valued objects exist in the given dataset, single-valued methods simply concatenate and regard the values provided by the same source as a single joint value. Specifically, given a multi-valued object $o$ ($|V_{s_o}| > 1$), they regard $V_{s_o}$ as a single joint value, denoted as $\mathcal{V}$, instead of considering each claimed value $v \in V_{s_o}$ individually. They label the values in $V_{s_o}$ as all true (i.e., $\mathcal{V}$ is true) or all false (i.e., $\mathcal{V}$ is false) together. In contrast, multi-valued methods treat the claimed values in $V_{s_o}$ individually and might assign different truth labels to the claimed values in each triple. Table 1 summarizes the notations used in this article. Due to the complexity of source trustworthiness calculation in the multi-valued scenario and the lack of synthetic dataset generator that generates datasets with multi-valued objects and complete ground truth, we leave the method comparison without using ground truth for multi-valued truth discovery (MTD) methods as our future work.

Formally, we name the actual value of an object $o$ the *ground truth* of $o$ (denoted by $V_o^*$), and the triple involves $o$ with the label "true" output by a truth discovery method $m$ the *identified truth* of $o$ (denoted by $V_o^m$). In single-valued scenario, $|V_o^*| = 1$, $|V_o^m| = 1$, while in multi-valued scenario, $|V_o^*|$, $|V_o^m|$ both might be greater than 1. After applying a group of truth discovery methods $M$ one-by-one on the triples, each method $m \in M$ outputs the *identified truth* for each object $o \in O$. The closer $V_o^m$ is to $V_o^*$ for each object, the better the method $m$ performs. We denote the *identified truth* of all objects in $O$ output by method $m$ as $V^m$ ($V_o^m \subset V^m$), and the *ground truth* of all objects in $O$, i.e., the *complete ground truth* of the given dataset, as $V^*$ ($V_o^* \subset V^*$). In most cases, the ground truth provided with each frequently utilized real-world dataset, denoted by $V^i$, is only a subset of the complete ground truth ($V^i \subset V^*$). We define the *coverage* of the ground truth as follows:

*Definition 3.1.* **Coverage of the Ground Truth** indicates the percentage of objects covered by the ground truth over all the objects in the given dataset. The coverage of the complete ground truth is 100%.

Given the output of each truth discovery method, i.e., $V^m$, $m \in M$, and the ground truth ($V^i$), the traditional ground truth–based evaluation approaches evaluate the effectiveness of each method in terms of *precision, recall, F1 score, accuracy/error Rate*, and *specificity* for categorical data. For each metric, the higher the value is, the better the method performs. In particular, to derive those five metrics, the ground truth–based approaches first produce a confusion matrix for each method. It cumulatively counts the numbers of true positives, false positives, true negatives, and false negatives for each object $o$ covered by $V^i$. Then, based on the matrix, it calculates the metrics. However, as $V^i$ is generally only a very small part of $V^*$, the distributions of true positives, false positives, true negatives, and false negatives obtained in this small sample space cannot reflect the real distributions. Therefore, the derived metrics are not statistically significant to be legitimately leveraged for method accuracy evaluation and comparison.

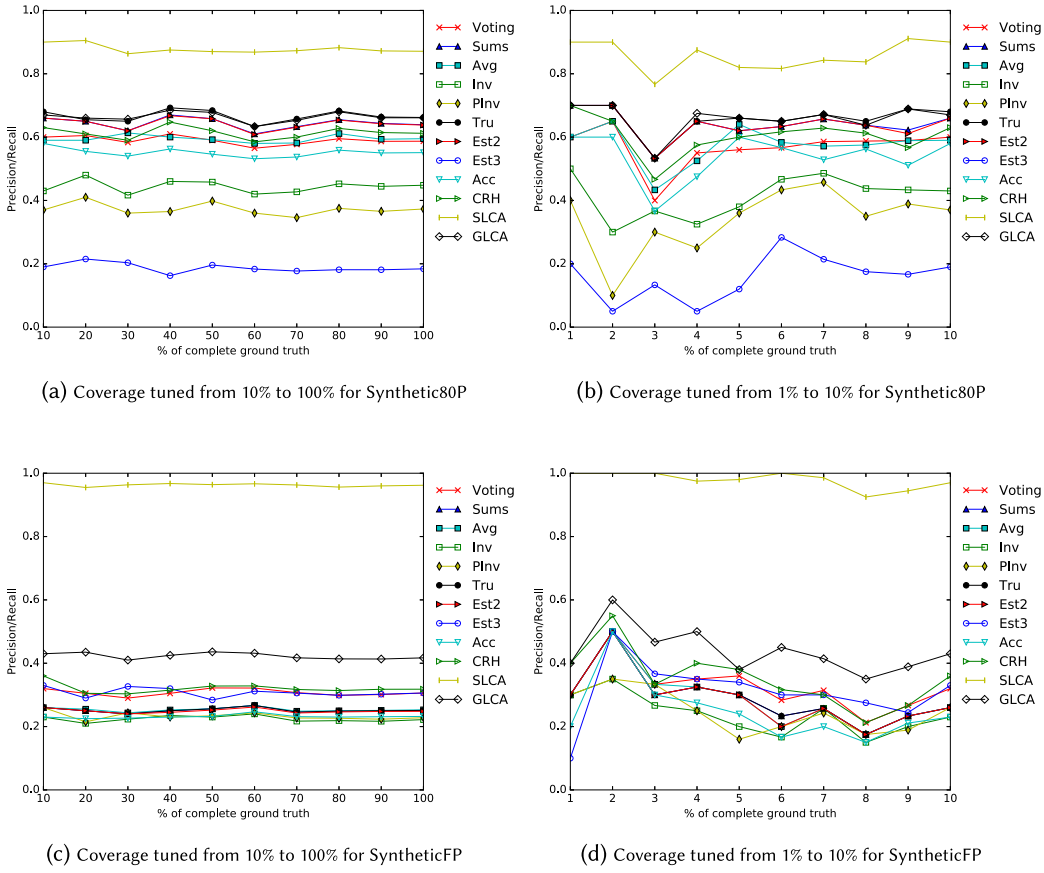
### 3.2 Motivation

As analyzed in Section 2, a range of truth discovery methods is proposed for different application scenarios. To include more methods in a comparable environment and make the computation of our approach tractable, we make the following assumptions:

- *Assumption 1 (Source consistency).* A source is likely to provide true information with the same probability for all the objects.
- *Assumption 2 (Source independence).* Data sources are independent of each other, i.e., no source-to-source relationship exists in the given dataset.
- *Assumption 3 (Object independence).* Objects are independent of each other, i.e., no object-to-object relationship exists in the given dataset.

We focus on categorical data type and single-valued truth discovery methods in this article and leave the fine-grained source reliability and enriched meaning of source reliability as our future work. As these 12 truth discovery methods, i.e., Majority voting, Accu [6],[2] TruthFinder [45], Sums, Average-Log, Investment, PooledInvestment [31, 32], 2-Estimates, 3-Estimates [16], SimpleLCA, GuessLCA [33], and CRH [21] are all *single-valued methods* that are compliant with the above assumptions and applicable for categorical data, we implement and compare them to describe the motivation of our work and evaluate our approach. In this section, we conduct empirical investigations on these 12 truth discovery methods using synthetic datasets with varied *coverages* of the ground truth to investigate the bias introduced by incomplete ground truth.

[2]Note that Accu is not compliant with Assumption 2, since it takes the copying relationship among sources into consideration. We chose this method to test how our approach performs when source-to-source copying relationship exists.




(a) Coverage tuned from 10% to 100% for Synthetic80P

(b) Coverage tuned from 1% to 10% for Synthetic80P

(c) Coverage tuned from 10% to 100% for SyntheticFP

(d) Coverage tuned from 1% to 10% for SyntheticFP

Fig. 1. Precision/Recall of 12 truth discovery methods evaluated on different coverages of the leveraged ground truth

The synthetic datasets with complete ground truth are generated by the dataset generator implemented by Waguih et al. [37]. This generator involves six parameters that are required to be configured to simulate a wide spectrum of truth discovery scenarios. We will introduce the settings of those parameters in detail in Section 5.1. We tuned the ground truth distribution per source (*GT*) for all the seven possible distributions, including *uniform, Random, Full-Pessimistic, Full-Optimistic, 80-Pessimistic, 80-Optimistic*, and *Exponential*. Based on the above configurations, we obtained seven dataset groups, each group containing 10 datasets. The metrics, namely, *precision, recall, F1 score, accuracy,* and *specificity* of each method were measured as the average of 10 executions over the 10 datasets included by the same dataset group. To calculate the metrics, for each dataset, we tuned the coverage of the ground truth from 10% to 100%, and also from 1% to 10%, by randomly picking up the specific quantity of objects from the complete ground truth. Due to the limited space, we only show the experimental results on two settings, namely, *80-Pessimistic* and *Full-Pessimistic*, with the corresponding datasets depicted as *Synthetic80P* and *SyntheticFP*. The experimental results on all the other datasets show the same results. Note that all the objects in the synthetic datasets have only one true value, thus the resulting precision, recall, and F1 score equal to each other. The accuracy and specificity show the same ranking results. Figure 1(a) and Figure 1(c) show the precision and recall of all the 12 methods with the coverage of the leveraged

ground truth tuned from 10% to 100%, while Figure 1(b) and Figure 1(d) show those of the methods with the coverage tuned from 1% to 10%. The latter range forms the sparse ground truth, which is closer to the reality, where the coverage of the collected ground truth is always below 10%, sometimes even below 1%.

Ideally, if the performance evaluation is not biased by the incomplete ground truth, there should be no intersecting lines in the figures, demonstrating that the ranking of the metrics of the methods is consistent with the results measured on complete ground truth. Even if two or more methods show the same performance, the precision/recall lines of those methods in the figures should totally overlap rather than intersect.

However, for both types of datasets, we cannot get the completely correct ranking for each type of datasets until the coverage of the leveraged ground truth grows up to 60%, which is generally impossible to obtain in reality. The results are even worse for the sparse ground truth. As shown in Figure 1(b) and Figure 1(d), by tuning the coverage of the ground truth, the ranking of methods fluctuates all the time, and no correct result is returned. That means the performance evaluation is strongly biased by the sparse ground truth. In most cases, real-world datasets would not have strict mathematical distributions, such as source coverage distributions, ground truth distribution per source, and distinct value distribution per object might be random. Therefore, the ranking based on real-world datasets with sparse ground truth would be even less correct.

## 4 OUR APPROACH

Under the *single-valued assumption*, by identifying a value of an object to be true, a truth discovery method is implying that all the other values of the object are false. When a method incorrectly identifies a false value of an object to be true, it certainly asserts the true value as a false value. In this case, the false positives are equivalent to false negatives, and the recall and F1 scores equal to the precision.

However, when it comes to the case of multi-valued objects, the identified truth of a multi-valued object may overlap with the ground truth. Simply labeling a value set as true or false according to whether it equals to the ground truth will degrade the accuracy of the performance evaluation of the method. For example, if the identified truth for "Tom's children" is {"Anna, Tim"}, and the ground truth is {"Anna, Tim, Lucas"}, the identified truth is partially true, rather than false. Therefore, we propose to treat each value in the identified value set individually. In this case, the false positives are no longer equivalent to false negatives. Given that neither the precision nor the recall of a method can reflect the performance of the method individually, we need to measure both the accuracy and the completeness of the methods' output. For example, given two methods $m_1$ and $m_2$, $m_1$ identifies {"Anna, Tim"} as "Tom's children," while $m_2$ identifies {"Anna"} is the only child of "Tom." The precision of both methods is 1, as their identified values are all true values, indicating their performances are the same. However, the recall of $m_1$ is $\frac{2}{3}$ and that of $m_2$ is $\frac{1}{3}$, indicating the performance of $m_1$ is better than $m_2$.

In this article, we evaluate the performance of methods separately for *single-valued scenarios* (i.e., scenarios where only single-valued objects exist) and *multi-valued scenarios* (i.e., scenarios where multi-valued objects exist). Note that the selected 12 truth discovery methods all make the single-valued assumption, but our comparison approach considers the multi-valued scenarios to evaluate those methods more accurately.

### 4.1 CompTruthHyp

The most straightforward approach for truth discovery is to conduct majority voting for categorical data or to average for continuous data. The largest limitation of such an approach is that it assumes all the sources are equally reliable, which does not hold in most real-world scenarios.

Thus, the most important feature of the existing truth discovery methods is their ability to estimate source trustworthiness [24]. While identifying the truth, current methods also return $c_{\mathcal{V}}$, the confidence score of each value $\mathcal{V}$ (or the probability of $\mathcal{V}$ being true), and $\tau_s$, the trustworthiness of each source $s$ (or the probability of source $s$ providing true information), as the intermediate variables. In particular, a higher $c_{\mathcal{V}}$ indicates that value $\mathcal{V}$ is more likely to be true, and a higher $\tau_s$ indicates that source $s$ is more reliable and the values claimed by this source are more likely to be true. Though the calculations of $c_{\mathcal{V}}$ and $\tau_s$ differ from one method to the other, current methods generally apply the same principle for truth discovery: If a source claims true values frequently, it will receive high trustworthiness; meanwhile, if a value is claimed by sources with high trustworthiness, it will be more likely to be identified as truth. To determine the truth, a weighted aggregation of the multi-source data is performed based on the estimated source trustworthiness. Thus, value confidence score and source trustworthiness calculation are the key elements for truth discovery and can be leveraged to compare the performance of current truth discovery methods.

In particular, we consider the output of each method, including value confidence score and source trustworthiness, as the hypotheses about the ground truth. The closer the hypotheses are to the ground truth, the better the method performs. As different method applies different model to estimate value confidence score and source trustworthiness, those measurements are incomparable between different methods. In our approach, we use value binary (i.e., true/false) labels instead of value confidence scores. We also unify source trustworthiness based on the value binary labels. Due to the lack of the ground truth, we take the observation of the dataset, or data distributions over sources, as the gold standard. After this transformation, the comparison of the performance of truth discovery methods becomes the comparison of their ability to infer the observation of the given dataset from their hypotheses. We fit the results of different methods into the data distribution of the given observation to see what is the resulting likelihood of observation conditioned on the hypotheses. The bigger the likelihood is, the better the method performs.

In this section, we present our approach, *CompTruthHyp*, which compares the 12 single-valued truth discovery methods without using ground truth in both single-valued scenarios and multi-valued scenarios. Our data model includes the following inputs: (i) the input dataset for truth discovery (i.e., $\{S, O, V\}$ triples); (ii) the identified truth of each method ($m \in M$, $|M| = 12$); (iii) source trustworthiness and value confidence scores output by each method. The output of our data model is a ranking of the accuracy of the 12 methods. As we do not have any ground truth, we propose to obtain the ranking by comparing the methods' ability to infer the observation of the given dataset from their outputs. We denote by $\phi$ the observation of which source votes for which value in the dataset, $V^m$ the identified truth of a method $m$, and $P(\phi|V^m)$ the probability of $\phi$ conditioned on $V^m$. A higher $P(\phi|V^m)$ indicates that the method $m$ has bigger ability to capture the features of the given dataset; thus, its output is more reliable.

Our computation requires several parameters, which can be derived from the inputs: $\tau_s(m)$, the probability that the claimed value of $s$ is true, given $V^m$. We will introduce the calculation of $\tau_s(m)$ in Section 4.2; $P_s(v_t|V_o^m)$ (respectively, $P_s(v_f|V_o^m)$), the probability that a source provides a particular true (respectively, false) value for object $o$, given $V_o^m$. We will introduce the calculations of $P_s(v_t|V_o^m)$ and $P_s(v_f|V_o^m)$ in Section 4.3. We compute the required parameters by applying different algorithms for single-valued and multi-valued scenarios.

In single-valued scenario, $|V_{s_o}| = 1$, $|V_o{}^m| = 1$. In multi-valued scenario, before applying our approach, we pre-process the given dataset by splitting each triple into $|V_{s_o}|$ triples by treating each claimed value individually. For example, a source $s$ claims "Tom's children" are {"Anna, Tim"}. The original triple is denoted as {$s$, "Tom's children", {"Anna, Tim"}}. After pre-processing, we get two triples, {$s$, "Tom's children," "Anna"} and {$s$, "Tom's children," "Tim"}. "Anna" and "Tim" are two claimed values by $s$ on object "Tom's children." Thus, in single-valued scenario, a claimed value

$v$ is equivalent to $V_{s_o}$. In the multi-valued scenario, a claimed value $v$ is a value in the value set $V_{s_o}$ claimed by a source on an object. Given $V^m$, ${V_o}^m \subset V^m$, $v$ is a claimed value, $V_o$ is the set of all claimed values on object $o$, if $v \in V_o^m$, then $v$ is identified as a true value by method $m$; if $v \in V_o - V_o^m$, then $v$ is identified as a false value by method $m$. Formally, if a source $s$ covers an object $o$, then we have the probability of the observation of $s$ providing a particular value $v$ ($v \in V_o$), conditioned on $V^m$, as:

$$P(\phi_{s_v}|V^m) = \begin{cases} \tau_s(m)P_s(v_t|V_o^m); & if\ v \in V_o^m \\ (1-\tau_s(m))P_s(v_f|V_o^m); & if\ v \in V_o - V_o^m \end{cases}. \tag{1}$$

In our observation, we are interested in two sets of values: given $V^m$, ${V_s}^t(m)$, denoting the set of true values provided by $s$; ${V_s}^f(m)$, denoting the set of false values provided by $s$. ${V_s}^t(m) \cup {V_s}^f(m) = V_s$, $V_s$ is the set of all values provided by $s$. Since we assume each source provides each value independently, we have the probability of the observation of source $s$ with its claimed values, i.e., $\phi_s$, conditioned on $V^m$, as:

$$P(\phi_s|V^m) = \left( \prod_{v \in {V_s}^t(m), o \in O} \tau_s(m)P_s(v_t|V_o^m) \prod_{v \in {V_s}^f(m), o \in O} (1-\tau_s(m))P_s(v_f|V_o^m) \right). \tag{2}$$

By assuming sources are independent on each other, the conditional probability of observing the given dataset $\phi$ is:

$$P(\phi|V^m) = \prod_{s \in S} \left( \prod_{v \in {V_s}^t(m), o \in O} \tau_s(m)P_s(v_t|V_o^m) \prod_{v \in {V_s}^f(m), o \in O} (1-\tau_s(m))P_s(v_f|V_o^m) \right). \tag{3}$$

To simplify the computation, we define the *confidence* of method $m$, denote by $C_m$, as

$$C_m = \sum_{s \in S} \left( \sum_{v \in {V_s}^t(m), o \in O} \ln \tau_s(m)P_s(v_t|V_o^m) + \sum_{v \in {V_s}^f(m), o \in O} \ln(1-\tau_s(m))P_s(v_f|V_o^m) \right). \tag{4}$$

### 4.2 Source Trustworthiness Normalization

The accuracy of truth discovery methods significantly depends on their source trustworthiness estimation. Although all methods calculate source trustworthiness as the weighted aggregation of value confidence scores, they adopt different models and equations. Therefore, the calculated

**ALGORITHM 1:** The algorithm of source trustworthiness normalization for the single-valued scenario

**Input**: Given dataset $\{S, O, V\}$ and $V^m$ for each $m \in M$.
**Output**: $\tau_s(m)$ for each $s \in S$, $m \in M$.

```
foreach m ∈ M do
    foreach s ∈ S do
        TP_s^m = 0; FP_s^m = 0;
        foreach o ∈ O_s do
            if V_{s_o} = V_o^m then
                TP_s^m ++;
            else
                FP_s^m ++;
        Calculate τ_s(m) by applying Equation (5);
return τ_s(m) for each s ∈ S, m ∈ M.
```

Table 2. Confusion Matrix of a Truth Discovery Method $m$

| | | Ground Truth | |
|---|---|---|---|
| | | True | False |
| Method $m$ | True | True Positive ($TP_m$) | False Positive ($FP_m$) |
| | False | False Negative ($FN_m$) | True Negative ($TN_m$) |

**ALGORITHM 2:** The algorithm of source trustworthiness normalization for the multi-valued scenario

**Input**: Given dataset $\{S, O, V\}$ and $V^m$ for each $m \in M$.
**Output**: $\tau_s(m)$ for each $s \in S$, $m \in M$.

```
1  foreach m ∈ M do
2      foreach s ∈ S do
3          TP_s^m = 0; FP_s^m = 0;
4          foreach o ∈ O do
5              foreach v ∈ V_{s_o} do
6                  if v ∈ V_o^m then
7                      TP_s^m ++;
8                  else
9                      FP_s^m ++;
10         Calculate τ_s(m) by applying Equation (5);
11 return τ_s(m) for each s ∈ S, m ∈ M.
```

source trustworthiness by each method has different meaning and is incomparable. To normalize source trustworthiness output by 12 methods, our approach, *CompTruthHyp*, regards the trustworthiness of a source as the probability of its claimed values being true (i.e., precision). We can derive a confusion matrix as shown in Table 2 for each source based on the identified truth of each method. Then, we calculate the precision of each source output by each method ($\tau_s(m)$) as follows:

$$\tau_s(m) = \frac{TP_s{}^m}{TP_s{}^m + FP_s{}^m}, \tag{5}$$

where $TP_s{}^m$ (respectively, $FP_s{}^m$) is the number of true positives (respectively, false positives) of the values claimed by source $s$, given $V^m$.

In the single-valued scenario, each source provides one value for any object of interest. Given $V^m$, all the values in $V_o - V_o{}^m$ are regarded as false ($|V_o - V_o{}^m| = |V_o| - 1$). We calculate $\tau_s(m)$ for each source by performing Algorithm 1. In particular, for each method $m \in M$ (Line 1), for each $s \in S$ (Line 2), for each $o \in O_S$ (Line 4, where $O_s$ is the objects covered by $s$), if $V_{s_o}$ is true (Line 5), then $TP_s{}^m$ increases by one (Line 6), otherwise, $FP_s{}^m$ increases by one (Lines 7, 8). For each source $s$, $\tau_s(m)$ is calculated by applying Equation (5) (Line 9).

In the multi-valued scenario, we calculate $\tau_s(m)$ for each source using Algorithm 2. As mentioned, $|V_o{}^m|$ and $|V_{s_o}|$ may be bigger than 1. Therefore, we first pre-process the dataset by splitting the triples and treating each claimed value in $V_{s_o}$ individually (Line 5).

### 4.3 True-false Distributions

Given the identified truth output by a truth discovery method, we analyze the true-false distribution of values for each object in the given dataset to calculate the probability that a source provides a particular true (respectively, false) value for an object.

For the single-valued scenario, each object has one single value. Therefore, we have $P_s(v_t|V_o^m)$ fixed to 1. As the false values may have varied distributions on an object, $P_s(v_f|V_o^m)$ can be different for each observed false value. Given a set of false values of $o$, $(V_o - V_o^m)$, we need to analyze their distribution and calculate the probability $(P_s(v_f|V_o^m))$ for sources to pick a particular value from the distribution. We define the *untrustworthiness* of a source as the probability that its claimed values are false, i.e., $(1 - \tau_s(m))$. For each particular false value $v_f$, each source that claims this value gives a vote of $(1 - \tau_s(m))$ for it being false. We consider there is a box containing all the false claims provided by all the sources in the given dataset. In this case, a particular claim $v_f$ may have several occurrences in the box if it is claimed by multiple sources $s \in S_{v_f}$. We count the occurrences of $v_f$ by $\sum_{s \in S_{v_f}} (1 - \tau_s(m))$. We calculate the probability of a particular false value being picked, i.e., $P_s(v_f|V_o^m)$ by:

$$P_s(v_f|V_o^m) = \frac{\sum_{s \in S_{v_f}} (1 - \tau_s(m))}{\sum_{v_f' \in V_o - V_o^m} \sum_{s' \in S_{v_f'}} (1 - \tau_{s'}(m))}, \tag{6}$$

where $S_{v_f}$ is the set of sources provide $v_f$ on $o$. We calculate this probability for each false value of each object using Algorithm 3.

**ALGORITHM 3:** The algorithm of $P_s(v_f|V_o^m)$ calculation for the single-valued scenario

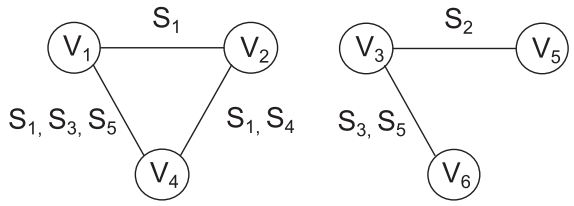
**Input**: Given dataset $\{S, O, V\}$ and $V^m$ for each $m \in M$.
**Output**: $P_s(v_f|V_o^m)$ for each $v_f \in V_o - V_o^m$, $o \in O$, $m \in M$.

```
foreach m ∈ M do
    foreach o ∈ O do
        foreach v_f ∈ V_o − V_o^m do
            foreach s ∈ S_{v_f} do
                P_s(v_f|V_o^m) += (1 − τ_s(m));
            P_s(v_f|V_o^m) of each v_f is normalized to satisfy Σ_{v_f ∈ V_o − V_o^m} P_s(v_f|V_o^m) = 1;
return P_s(v_f|V_o^m) for each v_f ∈ V_o − V_o^m, o ∈ O, m ∈ M.
```

In the multi-valued scenario, values in a source's claimed value set are not totally independent. Intuitively, the values occurring in the same claimed value set are believed to impact each other. The co-occurrence of values in the same claimed value set indicates that those values have potentially similar probabilities of being selected.

We define the weighted association among the distinctive values on the same object to represent their influence on each other, based on which to compute the probability of each value being selected. In particular, given $V_o^m$, we represent the bipartite mapping between true (respectively, false) values on each multi-valued object and sources that claim the true (respectively, false) values into a true (respectively, false) value graph. In each true (respectively, false) value graph, the identified true values (respectively, false values) in $V_o^m$ (respectively, $V_o - V_o^m$) are the vertices, and sources that claim those values are the weights of edges that connect with the values. For example, the value co-occurrences for a multi-valued object are shown in Figure 2. $V_o = \{v_1, v_2, v_3, v_4, v_5, v_6\}$, $V_o^m = \{v_1, v_2, v_4\}$. True values $v_2$ and $v_4$ are claimed by both $s_1$ and $s_4$, while false values $v_3$ and $v_5$ are claimed by $s_2$.

The detailed procedure of $P_s(v_t|V_o^m)$ and $P_s(v_f|V_o^m)$ calculation is shown in Algorithm 4. For each true (respectively, false) value graph, we further generate a corresponding square adjacent "true" (respectively, "false") matrix, which should be irreducible, aperiodic, and stochastic to be guaranteed to converge to a stationary state. In particular, we first initialize each element in the




(a) True value graph (b) False value graph

Fig. 2. An example of value co-occurrences for a multi-valued object.

matrix as the sum of the trustworthiness (respectively, untrustworthiness) of all sources that claim the co-occurrence of the corresponding pair of true (respectively, false) values (Line 8 and Line 17). To guarantee the three features of the matrix, we add a "*smoothing link*" by assigning a small weight to every pair of values (Line 9 and Line 18), where $\beta$ is the smoothing factor. For our experiments, we set $\beta = 0.1$ (empirical studies such as the work done by Gleich et al. [17] demonstrate more accurate estimation). We then normalize the elements to ensure that every column in the matrix sums to 1 (Line 10 and Line 19). This normalization allows us to interpret the elements as the transition probabilities for the random walk computation. Finally, we adopt the *Fixed Point Computation Model* (FPC) [3] on each "true" (respectively, "false") matrix to calculate $P_s(v_t|V_o^m)$ (respectively, $P_s(v_f|V_o^m)$) for each true (respectively, false) value of each object $o \in O$ (Line 11 and Line 20).

## 5 EXPERIMENTS

In this section, we first introduce our experimental setup in Section 5.1. Then, we report our evaluation results on both synthetic datasets in Section 5.2 and real-world datasets in Section 5.3.

### 5.1 Experimental Setup

*5.1.1 Evaluation Metrics.* We implemented all the 12 selected truth discovery methods, ground truth–based evaluation approach, and CompTruthHyp, in Python 3.4.0. All experiments were conducted on a 64-bit Windows 10 Pro. PC with an Intel Core i7-5600 processor and 16 GB RAM. We ran each truth discovery method 10 times and used the above-introduced five traditional evaluation metrics, including *precision, recall, accuracy, F1 score*, and *specificity*, as well as *confidence* output by CompTruthHyp, to evaluate their average performance. For the single-valued scenario, as the experimental results show that the rankings of different metrics are all equivalent, we discuss the precision of each method as an example. For the multi-valued scenario, we additionally introduce a new metric, namely, *average*, to measure the overall performance of the methods, which is calculated as the average of the precision, recall, accuracy, and specificity of each method.

To validate our approach, CompTruthHyp, we need to show the ranking of *confidence* of 12 selected methods is closer than the rankings of various evaluation metrics of the methods derived from sparse/low-quality ground truth to the real ranking of the performance of the methods derived from the complete ground truth. In this article, we adopt *Cosine similarity* (denoted as *Cos.*) and *Euclidean distance* (denoted as *Dist.*) to measure the distance of the two rankings. For Cosine similarity, a bigger value means better performance, while for Euclidean distance, a smaller value indicates better performance.

*5.1.2 Synthetic Datasets.* For the single-valued scenario, we applied the dataset generator introduced in Section 3.2, which can be configured to simulate a wide spectrum of truth discovery scenarios (except the multi-valued scenario). In particular, three parameters determine the scale

**ALGORITHM 4:** The algorithm of $P_s(v_t|V_o^m)$ and $P_s(v_f|V_o^m)$ calculation for the multi-valued scenario

**Input**: Given dataset $\{S, O, V\}$ and $V^m$ for each $m \in M$.
**Output**: $P_s(v_t|V_o^m)$ for each $v_t \in V_o^m$, $P_s(v_f|V_o^m)$ for each $v_f \in V_o - V_o^m$, $o \in O$, $m \in M$.

$\beta = 0.1$;
**foreach** $m \in M$ **do**
  // "true" matrix generation
  **foreach** $o \in O$ **do**
    **foreach** $v_{t_1} \in V_o^m$ **do**
      **foreach** $v_{t_2} \in V_o^m$ **do**
        **if** $v_{t_1} \neq v_{t_2}$ **then**
          **foreach** $s \in S_{v_{t_1}} \cap S_{v_{t_2}}$ **do**
            $TrueMatrix[v_{t_1}][v_{t_2}] += \tau_s(m)$;
          $TrueMatrix[v_{t_1}][v_{t_2}] = \beta + (1-\beta) * TrueMatrix[v_{t_1}][v_{t_2}]$;
    Normalize TrueMatrix;
    Apply FPC random walk computation to obtain $P_s(v_t|V_o^m)$ for each $v_t \in V_o^m$;
  // "false" matrix generation
  **foreach** $o \in O$ **do**
    **foreach** $v_{f_1} \in V_o - V_o^m$ **do**
      **foreach** $v_{f_2} \in V_o - V_o^m$ **do**
        **if** $v_{f_1} \neq v_{f_2}$ **then**
          **foreach** $s \in S_{v_{f_1}} \cap S_{v_{f_2}}$ **do**
            $FalseMatrix[v_{f_1}][v_{f_2}] += 1 - \tau_s(m)$;
          $FalseMatrix[v_{f_1}][v_{f_2}] = \beta + (1-\beta) * FalseMatrix[v_{f_1}][v_{f_2}]$;
    Normalize FalseMatrix;
    Apply FPC random walk computation to obtain $P_s(v_f|V_o^m)$ for each $v_f \in V_o - V_o^m$;
**return** $P_s(v_t|V_o^m)$ for each $v_t \in V_o^m$, $P_s(v_f|V_o^m)$ for each $v_f \in V_o - V_o^m$, $o \in O$, $m \in M$.

of the generated dataset, including the number of sources ($|S|$), the number of objects ($|O|$), and the number of distinct values per object ($|V_o|$). The other three parameters determine the characteristics of the generated dataset, including source coverage ($cov$), ground truth distribution per source ($GT$), and distinct value distribution per object ($conf$). We fixed the scale parameters by setting $|S| = 50$, $|O| = 1{,}000$, and $|V_o| = 20$. To better simulate the real-world scenarios, we configured both $cov$ and $conf$ as exponential distributions. By tuning $GT$ as all possible settings, including *uniform, Random, Full-Pessimistic, Full-Optimistic, 80-Pessimistic, 80-Optimistic*, and *Exponential*, we obtained eight groups of synthetic datasets (each group contains 10 datasets): (i) *U25* (Uniform 25), each source provides the same number (25%) of true positive claims; (ii) *U75* (Uniform 75), each source provides the same number (75%) of true positive claims; (iii) *80P* (80-Pessimistic), 80% of the sources provide 20% true positive claims; 20% of the sources provide 80% true positive claims. (iv) *80O* (80-Optimistic), 80% of the sources provide 80% true positive claims. 20% of the sources provide 20% true positive claims; (v) *FP* (Full-Pessimistic), 80% of the sources provide always false claims and 20% of the sources provide always true positive claims; (vi) *FO* (Full-Optimistic), 80% of the sources provide always true positive claims and 20% of the sources provide always false claims. (vii) *R* (Random), the number of true positive claims per source is

random; (viii) *Exp* (Exponential), the number of true positive values provided by the sources is exponentially distributed. All synthetic datasets were generated with the complete ground truth.

*5.1.3 Real-world Datasets.* We refined three real-world datasets for both single-valued and multi-valued scenarios: In particular, the Flight dataset, where each object only contains one true value, was applied for the single-valued scenario; and the Book-Author dataset and the Parent-Children dataset, where each object may contain multiple true values, were applied for the multi-valued scenario.

The *Flight* dataset was prepared by collecting gate information from the original Flight dataset [6]. The original Flight dataset contains 2,864,985 claims collected from 38 sources from the flight domain. The sources include 3 airline websites (AA, UA, Continental), 8 airport websites (such as SFO, DEN), and 27 third-party websites, including Orbitz, Travelocity, and so on. A claim represents the expected/actual departure/arrival time/gate of a particular flight on a particular day. It took the data provided by the three airline websites on 100 randomly selected flights as the gold standard. As this dataset is relatively big and our work focuses on categorical data, we refined and produced a new Flight dataset with complete ground truth by only reserving the departure/arrival gate of the flights covered in the original ground truth. The new dataset contains 38,493 distinctive claims provided by 21 sources.

The *Book-Author* dataset [45] contains 33,971 records crawled from *www.abebooks.com*. These records are collected from numerous book websites (i.e., sources). Each record represents a store's positive claims on the author list of a book (i.e., objects). 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 677 books, each book has on average 3 authors. The ground truth provided by the original dataset was utilized, which covers only 7.91% of the objects. The manually collected ground truth is sparse yet with high quality.

The *Parent-Children* dataset was prepared by extracting the parent-children relations from the *Biography dataset* [31]. We obtained 227,583 claims about the names of the children of 2,579 people (i.e., objects) edited by 54,764 users (i.e., sources). In the resulting dataset, each person has on average 2.48 children. We used the latest editing records as the ground truth, which covers all the objects. However, the quality of ground truth collected in this simple way is very poor.

### 5.2 Experiments on Synthetic Datasets

In this set of experiments, we aim to compare the confidence ($C_m$) and the precision of 12 methods calculated on different coverages of leveraged ground truth, denoted as P(1%) to P(100%), with their real precision calculated on the complete ground truth, denoted as P(100%), on eight groups of synthetic datasets with different settings of ground truth distributions. Tables 3 and 4 show the experimental results. As the results on U25 and U75 show similar features with 80P, we omit to show them in this article due to the limited space.

We observe that none of the 12 methods constantly outperforms the others in terms of precision, and a "one-fits-all" approach does not seem to be achievable. Based on the best performance values (shown in bold), we can see that the best method changed from dataset to dataset. In some cases, an improved method may not even beat its original version as a result of different features of the applied datasets. For example, while in most datasets 2-Estimates performed better than 3-Estimates, it performed worse than 3-Estimates in *FP* and *R*, where most of the claims provided by most sources could be false. This shows that in such cases, the factor that "hardness of facts"

Table 3. Experimental Results for Six Types of Representative Synthetic Datasets (the Single-valued Scenario)

| Dataset | Method | P(1%) | P(2%) | P(3%) | P(3%) | P(5%) | P(6%) | P(7%) | P(8%) | P(9%) | P(10%) | P(20%) | P(30%) | P(40%) | P(50%) | P(60%) | P(70%) | P(80%) | P(90%) | P(100%) | $C_m$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 80P | Voting | 0.600 | 0.650 | 0.400 | 0.550 | 0.560 | 0.567 | 0.586 | 0.588 | 0.589 | 0.600 | 0.605 | 0.583 | 0.610 | 0.590 | 0.565 | 0.577 | 0.595 | 0.587 | 0.587 | −16604 |
| | Sums | 0.700 | 0.700 | 0.533 | 0.650 | 0.620 | 0.633 | 0.657 | 0.638 | 0.622 | 0.660 | 0.650 | 0.620 | 0.670 | 0.658 | 0.610 | 0.633 | 0.655 | 0.643 | 0.639 | −16514 |
| | Avg | 0.600 | 0.650 | 0.433 | 0.525 | 0.640 | 0.583 | 0.571 | 0.575 | 0.589 | 0.590 | 0.590 | 0.613 | 0.600 | 0.592 | 0.580 | 0.581 | 0.611 | 0.593 | 0.595 | −16603 |
| | Inv | 0.500 | 0.300 | 0.367 | 0.325 | 0.380 | 0.467 | 0.486 | 0.438 | 0.433 | 0.430 | 0.480 | 0.417 | 0.460 | 0.458 | 0.420 | 0.427 | 0.453 | 0.444 | 0.448 | −17319 |
| | PInv | 0.400 | 0.100 | 0.300 | 0.250 | 0.360 | 0.433 | 0.457 | 0.350 | 0.389 | 0.370 | 0.410 | 0.360 | 0.365 | 0.398 | 0.360 | 0.346 | 0.375 | 0.366 | 0.373 | −17843 |
| | Tru | 0.700 | 0.700 | 0.533 | 0.650 | 0.660 | 0.650 | 0.671 | 0.650 | 0.689 | 0.680 | 0.655 | 0.650 | 0.693 | 0.684 | 0.633 | 0.657 | 0.683 | 0.663 | 0.662 | −16489 |
| | Est2 | 0.700 | 0.700 | 0.533 | 0.650 | 0.620 | 0.633 | 0.657 | 0.638 | 0.611 | 0.660 | 0.650 | 0.620 | 0.668 | 0.658 | 0.608 | 0.631 | 0.654 | 0.642 | 0.638 | −16514 |
| | Est3 | 0.200 | 0.05 | 0.133 | 0.050 | 0.120 | 0.283 | 0.214 | 0.175 | 0.167 | 0.190 | 0.215 | 0.203 | 0.163 | 0.196 | 0.183 | 0.177 | 0.181 | 0.181 | 0.184 | −18629 |
| | Accu | 0.600 | 0.600 | 0.367 | 0.475 | 0.600 | 0.567 | 0.529 | 0.563 | 0.511 | 0.580 | 0.555 | 0.540 | 0.563 | 0.546 | 0.532 | 0.537 | 0.559 | 0.550 | 0.551 | −16640 |
| | CRH | 0.700 | 0.650 | 0.467 | 0.575 | 0.600 | 0.617 | 0.629 | 0.613 | 0.567 | 0.630 | 0.610 | 0.590 | 0.648 | 0.620 | 0.585 | 0.600 | 0.628 | 0.614 | 0.612 | −16558 |
| | SLCA | **0.900** | **0.900** | **0.767** | **0.875** | **0.820** | **0.817** | **0.843** | **0.838** | **0.911** | **0.900** | **0.905** | **0.863** | **0.875** | **0.870** | **0.868** | **0.873** | **0.883** | **0.872** | **0.871** | **−15933** |
| | GLCA | 0.700 | 0.700 | 0.533 | 0.675 | 0.660 | 0.650 | 0.671 | 0.638 | 0.689 | 0.670 | 0.660 | 0.657 | 0.685 | 0.678 | 0.633 | 0.653 | 0.680 | 0.661 | 0.661 | −16480 |
| | **Dist.** | 5.916 | 4.359 | 3.873 | 3.000 | 4.123 | 1.732 | 2.000 | 2.646 | 3.162 | 1.732 | 2.236 | 2.236 | 1.414 | 1.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 1.414 |
| | **Cos.** | 0.975 | 0.987 | 0.989 | 0.993 | 0.987 | 0.998 | 0.997 | 0.995 | 0.992 | 0.998 | 0.996 | 0.996 | 0.998 | 0.999 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 | 0.998 |
| 80O | Voting | **1.000** | **0.950** | **1.000** | **1.000** | **1.000** | 0.983 | **1.000** | **1.000** | 0.989 | 0.990 | 0.985 | 0.990 | 0.983 | 0.992 | 0.992 | 0.989 | 0.991 | 0.991 | 0.992 | −10574 |
| | Sums | **1.000** | **0.950** | **1.000** | **1.000** | **1.000** | 0.983 | **1.000** | **1.000** | **1.000** | **1.000** | 0.985 | 0.990 | 0.985 | 0.992 | **0.993** | 0.991 | 0.994 | 0.993 | 0.994 | −10567 |
| | Avg | **1.000** | **0.950** | **1.000** | **1.000** | **1.000** | 0.983 | **1.000** | **1.000** | 0.989 | 0.990 | 0.985 | 0.990 | 0.983 | 0.992 | 0.992 | 0.989 | 0.991 | 0.991 | 0.992 | −10574 |
| | Inv | 0.900 | 0.800 | **1.000** | 0.875 | 0.800 | 0.817 | 0.871 | 0.900 | 0.889 | 0.900 | 0.815 | 0.833 | 0.813 | 0.834 | 0.837 | 0.834 | 0.859 | 0.844 | 0.847 | −11944 |
| | PInv | **1.000** | 0.850 | **1.000** | 0.925 | 0.940 | 0.850 | 0.914 | 0.950 | 0.944 | 0.960 | 0.880 | 0.897 | 0.868 | 0.888 | 0.892 | 0.890 | 0.905 | 0.893 | 0.898 | −11537 |
| | Tru | **1.000** | **0.950** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | 0.989 | 0.990 | 0.990 | **0.997** | 0.990 | 0.994 | 0.992 | 0.993 | 0.994 | 0.994 | 0.995 | −10554 |
| | Est2 | **1.000** | **0.950** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | 0.990 | **0.997** | 0.990 | 0.994 | **0.993** | **0.994** | **0.995** | **0.996** | **0.996** | −10554 |
| | Est3 | 0.200 | 0.150 | 0.200 | 0.125 | 0.200 | 0.150 | 0.257 | 0.138 | 0.133 | 0.250 | 0.155 | 0.210 | 0.195 | 0.192 | 0.212 | 0.184 | 0.206 | 0.193 | 0.196 | −12107 |
| | Accu | **1.000** | **0.950** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | 0.990 | **0.997** | 0.990 | 0.994 | **0.993** | **0.994** | **0.995** | **0.996** | **0.996** | −10554 |
| | CRH | **1.000** | **0.950** | **1.000** | **1.000** | **1.000** | 0.983 | **1.000** | **1.000** | 0.989 | **1.000** | 0.985 | 0.993 | 0.983 | 0.986 | 0.990 | 0.989 | 0.991 | 0.991 | 0.992 | −10577 |
| | SLCA | **1.000** | **0.950** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | 0.989 | 0.990 | **0.995** | **0.997** | **0.993** | **0.996** | **0.993** | **0.994** | **0.995** | **0.996** | **0.996** | **−10552** |
| | GLCA | **1.000** | **0.950** | **1.000** | **1.000** | **1.000** | 0.967 | **1.000** | **1.000** | 0.978 | 0.990 | 0.980 | 0.990 | 0.980 | 0.990 | 0.990 | 0.987 | 0.990 | 0.990 | 0.991 | −10578 |
| | **Dist.** | 15.427 | 12.530 | 12.530 | 12.530 | 12.530 | 3.610 | 12.530 | 12.530 | 5.477 | 8.426 | 3.162 | 4.796 | 2.646 | 4.359 | 4.123 | 1.000 | 0.000 | 1.000 | 0.000 | 3.317 |
| | **Cos.** | 0.778 | 0.859 | 0.859 | 0.859 | 0.859 | 0.989 | 0.859 | 0.859 | 0.975 | 0.938 | 0.992 | 0.981 | 0.994 | 0.984 | 0.985 | 0.999 | 1.000 | 0.999 | 1.000 | 0.991 |
| FP | Voting | 0.300 | 0.500 | 0.333 | 0.350 | 0.360 | 0.283 | 0.314 | 0.213 | 0.267 | 0.320 | 0.305 | 0.290 | 0.305 | 0.322 | 0.322 | 0.307 | 0.299 | 0.301 | 0.306 | −19758 |
| | Sums | 0.300 | 0.500 | 0.300 | 0.325 | 0.300 | 0.233 | 0.257 | 0.175 | 0.233 | 0.260 | 0.250 | 0.240 | 0.250 | 0.256 | 0.267 | 0.246 | 0.249 | 0.250 | 0.251 | −19647 |
| | Avg | 0.300 | 0.500 | 0.333 | 0.325 | 0.300 | 0.200 | 0.257 | 0.175 | 0.233 | 0.260 | 0.255 | 0.243 | 0.253 | 0.256 | 0.267 | 0.249 | 0.250 | 0.251 | 0.253 | −19660 |
| | Inv | 0.300 | 0.350 | 0.267 | 0.250 | 0.200 | 0.167 | 0.257 | 0.150 | 0.200 | 0.230 | 0.210 | 0.223 | 0.235 | 0.230 | 0.240 | 0.217 | 0.219 | 0.217 | 0.222 | −19841 |
| | PInv | 0.300 | 0.350 | 0.333 | 0.250 | 0.160 | 0.200 | 0.243 | 0.175 | 0.189 | 0.260 | 0.215 | 0.243 | 0.230 | 0.234 | 0.243 | 0.227 | 0.226 | 0.223 | 0.228 | −19931 |
| | Tru | 0.300 | 0.500 | 0.300 | 0.325 | 0.300 | 0.233 | 0.257 | 0.175 | 0.233 | 0.260 | 0.250 | 0.240 | 0.250 | 0.256 | 0.267 | 0.246 | 0.249 | 0.250 | 0.251 | −19639 |
| | Est2 | 0.300 | 0.500 | 0.300 | 0.325 | 0.300 | 0.200 | 0.257 | 0.175 | 0.233 | 0.260 | 0.250 | 0.240 | 0.245 | 0.252 | 0.262 | 0.243 | 0.246 | 0.248 | 0.248 | −19634 |
| | Est3 | 0.100 | 0.500 | 0.367 | 0.350 | 0.340 | 0.300 | 0.300 | 0.275 | 0.244 | 0.330 | 0.290 | 0.327 | 0.320 | 0.284 | 0.312 | 0.306 | 0.299 | 0.302 | 0.306 | −19526 |
| | Accu | 0.200 | 0.500 | 0.300 | 0.275 | 0.240 | 0.167 | 0.200 | 0.150 | 0.211 | 0.230 | 0.225 | 0.227 | 0.230 | 0.232 | 0.247 | 0.231 | 0.230 | 0.231 | 0.232 | −19576 |
| | CRH | 0.400 | 0.550 | 0.333 | 0.400 | 0.380 | 0.317 | 0.300 | 0.213 | 0.267 | 0.360 | 0.305 | 0.303 | 0.315 | 0.328 | 0.328 | 0.317 | 0.314 | 0.318 | 0.318 | −19703 |
| | SLCA | **1.000** | **1.000** | **1.000** | **0.975** | **0.980** | **1.000** | **0.986** | **0.925** | **0.944** | **0.970** | **0.955** | **0.963** | **0.968** | **0.964** | **0.967** | **0.963** | **0.956** | **0.960** | **0.962** | **−14860** |
| | GLCA | 0.400 | 0.600 | 0.467 | 0.500 | 0.380 | 0.450 | 0.414 | 0.350 | 0.389 | 0.430 | 0.435 | 0.410 | 0.425 | 0.436 | 0.432 | 0.417 | 0.414 | 0.413 | 0.417 | −19686 |
| | **Dist.** | 15.033 | 9.165 | 7.874 | 3.464 | 3.873 | 4.36 | 7.280 | 6.325 | 3.873 | 6.245 | 2.449 | 5.657 | 2.828 | 2.236 | 1.732 | 1.000 | 0.000 | 1.000 | 0.000 | 13.928 |
| | **Cos.** | 0.799 | 0.948 | 0.952 | 0.993 | 0.989 | 0.985 | 0.960 | 0.974 | 0.989 | 0.974 | 0.995 | 0.974 | 0.994 | 0.996 | 0.998 | 0.999 | 1.000 | 0.999 | 1.000 | 0.848 |

Table 4. Experimental Results for Six Types of Representative Synthetic Datasets (the Single-valued Scenario) Cont.

| Dataset | Method | P(1%) | P(2%) | P(3%) | P(3%) | P(5%) | P(6%) | P(7%) | P(8%) | P(9%) | P(10%) | P(20%) | P(30%) | P(40%) | P(50%) | P(60%) | P(70%) | P(80%) | P(90%) | P(100%) | $C_m$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FO | Voting | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **−1978** |
| | Sums | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **−1978** |
| | Avg | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **−1978** |
| | Inv | 0.500 | 0.350 | 0.500 | 0.650 | 0.380 | 0.400 | 0.571 | 0.550 | 0.578 | 0.490 | 0.490 | 0.450 | 0.458 | 0.460 | 0.478 | 0.447 | 0.456 | 0.477 | 0.464 | −10458 |
| | PInv | 0.500 | 0.350 | 0.400 | 0.575 | 0.340 | 0.317 | 0.529 | 0.525 | 0.500 | 0.420 | 0.420 | 0.407 | 0.410 | 0.412 | 0.425 | 0.397 | 0.409 | 0.422 | 0.412 | −11755 |
| | Tru | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **−1978** |
| | Est2 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **−1978** |
| | Est3 | 0.200 | 0.100 | 0.300 | 0.225 | 0.100 | 0.117 | 0.214 | 0.225 | 0.244 | 0.140 | 0.200 | 0.200 | 0.183 | 0.194 | 0.192 | 0.183 | 0.193 | 0.189 | 0.184 | −11885 |
| | Accu | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **−1978** |
| | CRH | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **−1978** |
| | SLCA | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **−1978** |
| | GLCA | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **−1978** |
| | **Dist.** | 1.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **Cos.** | 0.999 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| R | Voting | **0.400** | 0.200 | 0.233 | 0.250 | 0.320 | **0.350** | 0.300 | 0.288 | 0.267 | 0.340 | 0.340 | 0.313 | 0.290 | 0.306 | 0.283 | 0.301 | 0.314 | 0.310 | 0.303 | −19581 |
| | Sums | **0.400** | 0.150 | 0.267 | 0.250 | 0.320 | 0.333 | 0.286 | 0.288 | 0.267 | 0.360 | 0.335 | 0.317 | 0.270 | 0.306 | 0.293 | 0.299 | 0.310 | 0.307 | 0.303 | −19549 |
| | Avg | **0.400** | 0.150 | 0.300 | 0.250 | 0.320 | **0.350** | 0.343 | 0.263 | 0.278 | 0.350 | 0.315 | 0.323 | 0.300 | 0.310 | 0.293 | 0.307 | 0.320 | 0.311 | 0.307 | −19570 |
| | Inv | **0.400** | 0.200 | 0.300 | 0.275 | 0.300 | 0.333 | **0.371** | 0.263 | 0.289 | 0.370 | 0.325 | 0.337 | 0.303 | 0.292 | 0.300 | 0.313 | 0.321 | 0.319 | 0.312 | −19800 |
| | PInv | **0.400** | 0.200 | 0.267 | **0.325** | 0.300 | 0.333 | 0.329 | 0.238 | **0.311** | 0.340 | 0.340 | 0.330 | 0.305 | 0.292 | 0.308 | **0.320** | 0.324 | 0.323 | 0.317 | −19839 |
| | Tru | **0.400** | 0.150 | 0.233 | 0.225 | 0.360 | **0.350** | 0.271 | **0.300** | 0.267 | 0.350 | 0.315 | 0.323 | 0.293 | 0.310 | 0.293 | 0.301 | 0.315 | 0.311 | 0.306 | −19522 |
| | Est2 | **0.400** | 0.150 | 0.267 | 0.250 | 0.320 | 0.333 | 0.286 | 0.288 | 0.267 | 0.360 | 0.330 | 0.317 | 0.270 | 0.304 | 0.292 | 0.299 | 0.309 | 0.306 | 0.302 | −19548 |
| | Est3 | **0.400** | 0.200 | 0.333 | 0.175 | 0.300 | 0.250 | 0.214 | 0.275 | 0.300 | **0.380** | 0.330 | **0.370** | **0.348** | 0.320 | **0.327** | 0.313 | **0.333** | **0.332** | **0.330** | **−19356** |
| | Accu | **0.400** | 0.150 | 0.267 | 0.250 | 0.320 | 0.350 | 0.286 | 0.263 | 0.244 | 0.350 | 0.310 | 0.313 | 0.280 | 0.302 | 0.290 | 0.304 | 0.310 | 0.308 | 0.302 | −19551 |
| | CRH | **0.400** | **0.250** | 0.233 | 0.200 | 0.320 | **0.350** | 0.243 | 0.275 | 0.267 | 0.350 | **0.350** | 0.297 | 0.273 | 0.294 | 0.293 | 0.303 | 0.308 | 0.306 | 0.299 | −19496 |
| | SLCA | **0.400** | 0.200 | **0.367** | 0.250 | **0.380** | 0.300 | 0.271 | 0.275 | 0.244 | 0.370 | **0.350** | 0.330 | 0.315 | **0.334** | 0.300 | 0.304 | 0.313 | 0.317 | 0.313 | −19376 |
| | GLCA | **0.400** | 0.200 | 0.167 | 0.250 | 0.360 | **0.350** | 0.286 | 0.288 | 0.267 | 0.350 | 0.325 | 0.323 | 0.295 | 0.314 | 0.297 | 0.304 | 0.309 | 0.309 | 0.305 | −19531 |
| | **Dist.** | 21.726 | 14.457 | 14.318 | 17.635 | 19.875 | 23.601 | 16.279 | 19.950 | 12.884 | 14.177 | 16.217 | 4.243 | 4.359 | 13.153 | 7.937 | 8.718 | 6.000 | 3.464 | 0.000 | 16.733 |
| | **Cos.** | 0.887 | 0.814 | 0.818 | 0.713 | 0.625 | 0.488 | 0.774 | 0.648 | 0.854 | 0.821 | 0.781 | 0.985 | 0.985 | 0.858 | 0.947 | 0.936 | 0.971 | 0.990 | 1.000 | 0.778 |
| Exp | Voting | 0.000 | 0.000 | 0.267 | 0.175 | 0.120 | 0.167 | 0.186 | 0.175 | 0.167 | 0.190 | 0.145 | 0.153 | 0.155 | 0.150 | 0.145 | 0.136 | 0.146 | 0.151 | 0.145 | −19530 |
| | Sums | 0.000 | 0.000 | 0.267 | 0.175 | 0.120 | 0.167 | 0.186 | 0.175 | 0.167 | 0.190 | 0.145 | 0.153 | 0.155 | 0.150 | 0.145 | 0.136 | 0.146 | 0.151 | 0.145 | −19488 |
| | Avg | 0.000 | 0.000 | 0.267 | 0.175 | 0.120 | 0.167 | 0.186 | 0.175 | 0.167 | 0.190 | 0.145 | 0.153 | 0.155 | 0.150 | 0.145 | 0.136 | 0.146 | 0.151 | 0.145 | −19508 |
| | Inv | 0.000 | 0.000 | 0.367 | 0.325 | 0.160 | 0.233 | 0.257 | 0.250 | 0.211 | 0.240 | 0.205 | 0.203 | 0.228 | 0.210 | 0.203 | 0.190 | 0.208 | 0.206 | 0.202 | −20000 |
| | PInv | 0.000 | 0.000 | 0.400 | 0.350 | 0.160 | 0.250 | 0.286 | 0.275 | 0.256 | 0.250 | 0.235 | 0.217 | 0.260 | 0.238 | 0.223 | 0.210 | 0.236 | 0.231 | 0.229 | −20185 |
| | Tru | 0.000 | 0.000 | 0.267 | 0.175 | 0.120 | 0.167 | 0.186 | 0.175 | 0.167 | 0.190 | 0.145 | 0.153 | 0.155 | 0.150 | 0.145 | 0.136 | 0.146 | 0.151 | 0.145 | −19474 |
| | Est2 | 0.000 | 0.000 | 0.267 | 0.175 | 0.120 | 0.167 | 0.186 | 0.175 | 0.167 | 0.190 | 0.145 | 0.153 | 0.155 | 0.15 | 0.145 | 0.136 | 0.146 | 0.151 | 0.145 | −19488 |
| | Est3 | **0.420** | **0.420** | **0.420** | **0.420** | **0.420** | **0.420** | **0.420** | **0.420** | **0.420** | **0.420** | **0.420** | **0.420** | **0.420** | **0.420** | **0.420** | **0.420** | **0.420** | **0.420** | **0.420** | **−16596** |
| | Accu | 0.000 | 0.000 | 0.267 | 0.175 | 0.120 | 0.167 | 0.186 | 0.175 | 0.167 | 0.190 | 0.145 | 0.153 | 0.155 | 0.150 | 0.145 | 0.136 | 0.146 | 0.151 | 0.145 | −19493 |
| | CRH | 0.000 | 0.000 | 0.267 | 0.175 | 0.120 | 0.167 | 0.186 | 0.175 | 0.167 | 0.190 | 0.145 | 0.153 | 0.155 | 0.150 | 0.145 | 0.136 | 0.146 | 0.151 | 0.145 | −19482 |
| | SLCA | 0.200 | 0.200 | 0.333 | 0.225 | 0.220 | 0.233 | 0.243 | 0.238 | 0.278 | 0.270 | 0.270 | 0.300 | 0.240 | 0.246 | 0.268 | 0.250 | 0.261 | 0.268 | 0.257 | −19362 |
| | GLCA | 0.000 | 0.000 | 0.300 | 0.200 | 0.120 | 0.183 | 0.186 | 0.188 | 0.178 | 0.200 | 0.160 | 0.163 | 0.168 | 0.162 | 0.155 | 0.146 | 0.155 | 0.160 | 0.153 | −19489 |
| | **Dist.** | 8.246 | 8.246 | 6.164 | 2.449 | 2.828 | 1.732 | 3.606 | 2.449 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 1.414 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 13.114 |
| | **Cos.** | 0.978 | 0.978 | 0.974 | 0.990 | 0.998 | 0.995 | 0.986 | 0.990 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.997 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.873 |

should be considered to achieve better truth discovery. This instability of truth discovery methods reveals the importance of evaluating the methods. With a better evaluation approach, users can choose the best method for truth discovery more easily and accurately for a given scenario.

From the table, we can see that CompTruthHyp can always identify the best method for the given dataset. For *80P*, *80O*, and *FO*, the majority of methods performed better than random guessing with the real precision bigger than 0.5. For *FO*, the ranking of precision stayed stable with the coverage of the ground truth tuned from 1% to 100% and was consistent with the ranking of the real precision. The ranking of the confidence of methods output by CompTruthHyp was also equal to the ranking of their real precision, with $Dist. = 0$ and $Cos. = 1$. While CompTruthHyp and ground truth–based evaluation approach showed similar performance on this type of dataset in terms of accuracy, our approach did not cost any efforts for ground truth collection. For *80P* and *80O*, when the coverage of the ground truth increased, the Euclidean distance decreased until it reached 0 (70% for *80P*, 80% for *80O*), the Cosine similarity increased until it reached 1 (70% for *80P*, 80% for *80O*). The Euclidean distance and Cosine similarity of the confidence ranking were 1.414 and 0.998 for *80P*, which were as good as those of P(40%), while for *80O*, the ground truth–based evaluation approach beat our approach only when they got a ground truth with coverage bigger than 70%. Moreover, in real-world datasets, the collection of a ground truth with coverage bigger than 10% is a rather challenging task.

For *R*, *FP*, and *Exp*, none of the methods was reliable, except for SLCA on *FP*. Almost all the methods performed worse than random guessing with a real precision smaller than 0.5, and the real precision of those methods was similar with each other. For *R*, with the coverage of the ground truth increased, the Euclidean distance and Cosine similarity of the precision ranking fluctuated.

Even when the coverage reached 90%, the Euclidean distance was 3.464, which is still not close enough to the real ranking. Though the Euclidean distance of the confidence ranking was 16.733 and the Cosine similarity was 0.778, which are not close to the real ranking, it performed better than the rankings of P(1%), P(4%), P(5%), P(6%), P(8%) in terms of Euclidean distance, and those of P(4%), P(5%), P(6%), P(7%), P(8%) in terms of Cosine similarity. In the case of *FP*, our approach can only identify the best method and performed better than the ground truth–based evaluation approach when the coverage of the ground truth was 1%. However, in this case, only the best method performed better than random guessing and all the other methods showed very similar bad performance. For *Exp*, where one source always lies and one source always tells the truth for all the objects and the remaining sources range from 1% to 99% of values they claim is true, none of the methods was reliable and all of them performed similarly bad. Even in this case, our approach can still find out the best method, i.e., 3-Estimates.

### 5.3 Experiments on Real-world Datasets

In this set of experiments, we report comparative studies with three real-world datasets. For the single-valued scenario, we applied the Flight dataset with complete ground truth. We aim to compare the ranking of confidence ($C_m$) of the 12 methods with the ranking of their real precision calculated on the complete ground truth. Table 5 shows the experimental results, with the top-three best performances in bold. As there are five groups of sources with potential copy in this dataset and only Accu took the copying relations into consideration, Accu performed the best among the 12 methods. Though the Euclidean distance of the confidence ranking is 6.164 and the Cosine similarity is 0.971, which are not perfectly close to the real ranking, our approach successfully labels the best two methods as well as the four worst methods. The reasons why our approach did not achieve the perfect ranking may include: (i) We observed data sharing between sources, and even on low-quality data in this dataset, and this violated the source independence assumption made by our approach. We will consider to relax this assumption in our future work. (ii) Generating gold

Table 5. Experimental Results for Flight Dataset with Complete Ground Truth (the Single-valued Scenario)

| Dataset | Method | Precision | $C_m$ |
|---|---|---|---|
| **Flight** | Voting | 0.889 | −21564 |
| | Sums | **0.915** | −20281 |
| | AvgLog | 0.914 | −20326 |
| | Inv | 0.636 | −28108 |
| | PInv | 0.691 | −27195 |
| | Tru | 0.818 | −22925 |
| | Est2 | 0.887 | −20065 |
| | Est3 | 0.562 | −29167 |
| | Accu | **0.940** | **−19738** |
| | CRH | **0.923** | **−19748** |
| | SLCA | 0.893 | **−19855** |
| | GLCA | 0.894 | −21419 |
| | Dist. | 0.000 | 6.164 |
| | Cos. | 1.000 | 0.971 |

standards is challenging when we cannot observe the real world in person but have to trust some particular sources. Since every source can make mistakes, the gold standard of the Flight dataset could contain errors.

For multi-valued scenario, as precision cannot reflect the overall performance of a method with the complete ground truth (as analyzed in Section 3.1), we compared the confidence ranking of the methods with the ranking of all six metrics calculated on the provided ground truth, including precision, recall, accuracy, specificity, F1 score, and average. Table 6 shows the experimental results, with the top-three best performances in bold. These results also validate the observation that no method constantly outperforms the others. We also observed that the rankings of different metrics differed from one another, which validates our assertion that any one of those metrics can not individually reflect the overall performance of the methods. All methods performed worse on the Book-Author dataset than on the Parent-Children dataset with lower precision, recall, accuracy, and specificity. The possible reasons contain the poorer quality of sources (poorer ground truth distribution), more missing values (i.e., true values that are missed by all the sources), and the smaller dataset size.

For both datasets, our approach can consistently identify the top-three best methods. The confidence ranking is more similar with the ranking of average than the ranking using other metrics. This validates that confidence metric reflects the overall performance of the methods. However, for the Book-Author dataset, the Euclidean distance of the confidence ranking to average was still bigger than 4.0, and the Cosine similarity with average was still lower than 0.99. This is because the ground truth is relatively sparse, so the ranking of average cannot reflect the real performance ranking of the methods. Another reason is that there may be copying relations among sources, which are neglected by all the methods including our approach. Compared with the Book-Author dataset, the confidence ranking was closer to the rankings of all metrics on the Parent-Children dataset. This is because the ground truth covers all the objects and is obtained by collecting all the latest editions regarding the objects. Although the precision of the ground truth does not reach 1,

Table 6. Experimental Results for Two Real-world Datasets (the Multi-valued Scenario)

| Dataset | Method | Precision | Recall | Accuracy | Specificity | F1 | Average | $C_m$ |
|---|---|---|---|---|---|---|---|---|
| **Book** | Voting | 0.749 | 0.712 | 0.576 | 0.022 | 0.730 | 0.515 | −26258 |
| | Sums | 0.851 | 0.685 | 0.651 | **0.511** | 0.759 | 0.674 | −23011 |
| | AvgLog | 0.841 | 0.663 | 0.629 | 0.489 | 0.742 | 0.656 | −23477 |
| | Inv | 0.815 | 0.745 | 0.659 | 0.311 | 0.778 | 0.633 | −23860 |
| | PInv | 0.812 | 0.750 | 0.659 | 0.289 | 0.780 | 0.628 | −23435 |
| | Tru | 0.847 | 0.663 | 0.633 | **0.511** | 0.744 | 0.664 | −23303 |
| | Est2 | **0.863** | **0.755** | **0.707** | **0.511** | **0.806** | **0.709** | **−21915** |
| | Est3 | 0.828 | 0.734 | 0.664 | 0.378 | 0.778 | 0.651 | −24907 |
| | Accu | **0.858** | **0.788** | **0.725** | 0.467 | **0.822** | **0.709** | **−21390** |
| | CRH | 0.850 | 0.679 | 0.646 | **0.511** | 0.755 | 0.672 | −22751 |
| | SLCA | **0.861** | **0.810** | **0.742** | 0.467 | **0.835** | **0.720** | **−21670** |
| | GLCA | 0.846 | 0.658 | 0.629 | **0.511** | 0.740 | 0.661 | −23243 |
| | Dist. | 5.099 | 13.153 | 11.225 | 13.153 | 10.863 | 4.472 | 0.000 |
| | Cos | 0.980 | 0.865 | 0.901 | 0.861 | 0.909 | 0.985 | 1.000 |
| **Parent** | Voting | 0.919 | 0.901 | 0.845 | 0.462 | 0.910 | 0.782 | −330234 |
| | Sums | 0.938 | 0.927 | 0.883 | 0.585 | 0.933 | 0.833 | −314582 |
| | AvgLog | 0.938 | 0.926 | 0.882 | 0.581 | 0.932 | 0.832 | −314124 |
| | Inv | 0.915 | 0.919 | 0.841 | 0.457 | 0.917 | 0.783 | −331351 |
| | PInv | 0.912 | 0.912 | 0.839 | 0.454 | 0.912 | 0.779 | −331523 |
| | Tru | 0.938 | 0.926 | 0.881 | 0.581 | 0.932 | 0.832 | −315231 |
| | Est2 | **0.940** | **0.927** | **0.885** | **0.595** | **0.933** | **0.836** | **−309873** |
| | Est3 | 0.905 | 0.889 | 0.822 | 0.366 | 0.897 | 0.746 | −340031 |
| | Accu | **0.941** | **0.928** | **0.885** | **0.588** | **0.934** | **0.836** | **−310314** |
| | CRH | 0.938 | 0.927 | 0.883 | 0.586 | 0.932 | 0.833 | −313421 |
| | SLCA | **0.942** | **0.927** | **0.886** | **0.601** | **0.935** | **0.839** | **−302873** |
| | GLCA | 0.938 | 0.924 | 0.876 | 0.578 | 0.931 | 0.829 | −321098 |
| | Dist. | 2.828 | 3.742 | 2.000 | 1.000 | 3.162 | 1.414 | 0.000 |
| | Cos. | 0.994 | 0.989 | 0.997 | 0.999 | 0.992 | 0.998 | 1.000 |

the quality of sources in this dataset is relatively high. Therefore, the leveraged ground truth is similar to the complete ground truth.

## 6 CONCLUSION

In this article, we focus on the problem of comparing truth discovery methods without using the ground truth, which has not been studied by previous research efforts. We first motivate this study by revealing the bias introduced by sparse ground truth in evaluating the truth discovery methods by conducting experiments on synthetic datasets with different coverages of the ground truth. Then, we propose a generic approach, called *CompTruthHyp*, to solve this bias. In particular, we propose two approaches for single-valued and multi-valued scenarios, respectively. Given a dataset, we first calculate the precision of each source by the output of each truth discovery method. Based on the source precision and the identified truth, we estimate the probability of observations of the given data set for each method. The performance of methods is determined by the ranking of the calculated probabilities. Experimental studies on both real-world and synthetic datasets demonstrate the effectiveness of our approach.

This article is our first step towards truth discovery methods comparison without using the ground truth. Our future work will focus on enhancing the approach by considering more complex application scenarios. For example, we are interested in the scenarios with complex source relationships such as copying and mutual supportive relations (i.e., two sources with similar facts) [24].

## REFERENCES


[1] Djamal Benslimane, Quan Z. Sheng, Mahmoud Barhamgi, and Henri Prade. 2016. The uncertain web: Concepts, challenges, and current solutions. *ACM Trans. Internet Technol.* 16, 1 (2016), 1:1–1:6.

[2] Laure Berti-Équille. 2019. *Truth Discovery*. Springer International Publishing, Cham, 1–8.

[3] Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Comput. Netw. ISDN Syst.* 30, 1–7 (1998), 107–117.

[4] Anish Das Sarma, Xin Dong, and Alon Halevy. 2011. Data integration with dependent sources. In *Proceedings of the 14th International Conference on Extending Database Technology (EDBT'11)*. 401–412.

[5] Xin Luna Dong, Laure Berti-Equille, Yifan Hu, and Divesh Srivastava. 2010. Global detection of complex copying relationships between sources. *Proc. VLDB Endow.* 3, 1–2 (2010), 1358–1369.

[6] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. 2009. Integrating conflicting data: The role of source dependence. *Proc. VLDB Endow.* 2, 1 (2009), 550–561.

[7] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. 2009. Truth discovery and copying detection in a dynamic world. *Proc. VLDB Endow.* 2, 1 (2009), 562–573.

[8] Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 601–610.

[9] Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Kevin Murphy, Shaohua Sun, and Wei Zhang. 2014. From data fusion to knowledge fusion. *Proc. VLDB Endow.* 7, 10 (2014), 881–892.

[10] Xin Luna Dong, Barna Saha, and Divesh Srivastava. 2012. Less is more: Selecting sources wisely for integration. *Proc. VLDB Endow.* 6, 2 (2012), 37–48.

[11] Wenfei Fan. 2012. Data quality: Theory and practice. In *Proceedings of the International Conference on Web-Age Information Management*. 1–16.

[12] Wenfei Fan, Floris Geerts, Shuai Ma, Nan Tang, and Wenyuan Yu. 2013. Data quality problems beyond consistency and duduplication. In *Search of Elegance in the Theory and Practice of Computation: Essays Dedicated to Peter Buneman*. Springer Berlin Heidelberg, 237–249.

[13] Xiu Susie Fang, Quan Z. Sheng, Xianzhi Wang, Mahmoud Barhamgi, Lina Yao, and Anne H. H. Ngu. 2017. SourceVote: Fusing multi-valued data via inter-source agreements. In *Proceedings of the 36th International Conference on Conceptual Modeling (ER'17)*. 164–172.

[14] Xiu Susie Fang, Quan Z. Sheng, Xianzhi Wang, Dianhui Chu, and Anne H. H. Ngu. 2019. SmartVote: A full-fledged graph-based model for multi-valued truth discovery. *World Wide Web J.* 22, 4 (2019), 1855–1885.

[15] Xiu Susie Fang, Quan Z. Sheng, Xianzhi Wang, and Anne H. H. Ngu. 2017. Value veracity estimation for multi-truth objects via a graph-based approach. In *Proceedings of the International World Wide Web Conference (WWW'17)*. 777–778.

[16] Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. 2010. Corroborating information from disagreeing views. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM'10)*. 131–140.

[17] David Gleich, Paul Constantine, Abraham Flaxman, and Asela Gunawardana. 2010. Tracking the random surfer: Empirically measured teleportation parameters in PageRank. In *Proceedings of the International World Wide Web Conference (WWW'10)*. 381–390.

[18] Manish Gupta, Yizhou Sun, and Jiawei Han. 2011. Trust analysis with clustering. In *Proceedings of the International World Wide Web Conference (WWW'11)*. 53–54.

[19] Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *J. ACM* 46, 5 (1999), 604–632.

[20] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. 2014. A confidence-aware approach for truth discovery on long-tail data. *Proc. VLDB Endow.* 8, 4 (2014).

[21] Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei Han. 2014. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 1187–1198.

[22] Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. 2012. Truth finding on the deep web: Is the problem solved? *Proc. VLDB Endow.* 6, 2 (2012), 97–108.

[23] Xian Li, Xin Luna Dong, Kenneth B. Lyons, Weiyi Meng, and Divesh Srivastava. 2015. Scaling up copy detection. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE'15)*. 89–100.

[24] Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2015. A survey on truth discovery. *ACM SIGKDD Explor. Newslett.* 17, 2 (2015), 1–16.

[25] Yaliang Li, Chenglin Miao, Lu Su, Jing Gao, Qi Li, Bolin Ding, Zhan Qin, and Kui Ren. 2018. An efficient two-layer mechanism for privacy-preserving truth discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'18)*. 1705–1714.

[26] Xueling Lin and Lei Chen. 2018. Domain-aware multi-truth discovery from conflicting sources. *Proc. VLDB Endow.* 11, 5 (2018), 635–647.

[27] Xuan Liu, Xin Luna Dong, Beng Chin Ooi, and Divesh Srivastava. 2011. Online data fusion. *Proc. VLDB Endow.* 4, 11 (2011), 932–943.

[28] J. Marshall, A. Argueta, and D. Wang. 2017. A neural network approach for truth discovery in social sensing. In *Proceedings of the IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS'17)*. 343–347.

[29] Chenglin Miao, Wenjun Jiang, Lu Su, Yaliang Li, Suxin Guo, Zhan Qin, Houping Xiao, Jing Gao, and Kui Ren. 2019. Privacy-preserving truth discovery in crowd sensing systems. *ACM Trans. Sens. Netw.* 15, 1 (2019).

[30] Jeff Pasternack and Dan Roth. 2010. Comprehensive trust metrics for information networks. In *Proceedings of the Army Science Conference*.

[31] Jeff Pasternack and Dan Roth. 2010. Knowing what to believe (when you already know something). In *Proceedings of the International Conference on Computational Linguistics (COLING'10)*. 877–885.

[32] Jeff Pasternack and Dan Roth. 2011. Making better informed trust decisions with generalized fact-finding. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI'11)*. 2324–2329.

[33] Jeff Pasternack and Dan Roth. 2013. Latent credibility analysis. In *Proceedings of the International World Wide Web Conference (WWW'13)*. 1009–1020.

[34] Ravali Pochampally, Anish Das Sarma, Xin Luna Dong, Alexandra Meliou, and Divesh Srivastava. 2014. Fusing data with correlations. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 433–444.

[35] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the International World Wide Web Conference (WWW'17)*. 1003–1012.

[36] Theodoros Rekatsinas, Xin Luna Dong, and Divesh Srivastava. 2014. Characterizing and selecting fresh data sources. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 919–930.

[37] Dalia Attia Waguih and Laure Berti-Equille. 2014. Truth discovery algorithms: An experimental evaluation. *Arxiv Preprint Arxiv:1409.6428* (2014).

[38] Mengting Wan, Xiangyu Chen, Lance Kaplan, Jiawei Han, Jing Gao, and Bo Zhao. 2016. From truth discovery to trustworthy opinion discovery: An uncertainty-aware quantitative modeling approach. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1885–1894.

[39] D. Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. C. Aggarwal, R. Ganti, X. Wang, P. Mohapatra, B. Szymanski, and H. Le. 2014. Using humans as sensors: An estimation-theoretic perspective. In *Proceedings of the of the International Conference on Information Processing in Sensor Networks (IPSN'14)*. 35–46.

[40] Xianzhi Wang, Quan Z. Sheng, Xiu Susie Fang, Lina Yao, Xiaofei Xu, and Xue Li. 2015. An integrated Bayesian approach for effective multi-truth discovery. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM'15)*. 493–502.

[41] Xianzhi Wang, Quan Z. Sheng, Lina Yao, Xiu Susie Fang, Xiaofei Xu, and Boualem Benatallah. 2016. Truth discovery via exploiting implications from multi-source data. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM'16)*. 861–870.

[42] Xianzhi Wang, Quan Z. Sheng, Lina Yao, Xue Li, Xiu Susie Fang, and Xiaofei Xu. 2016. Empowering truth discovery with multi-truth prediction. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM'16)*. 881–890.

[43] Houping Xiao, Jing Gao, Qi Li, Fenglong Ma, Lu Su, Yunlong Feng, and Aidong Zhang. 2016. Towards confidence in the truth: A bootstrapping based truth discovery approach. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1935–1944.

[44] H. Xiao, J. Gao, Q. Li, F. Ma, L. Su, Y. Feng, and A. Zhang. 2019. Towards confidence interval estimation in truth discovery. *IEEE Trans. Knowl. Data Eng.* 31, 3 (Mar. 2019), 575–588.

[45] Xiaoxin Yin, Jiawei Han, and Philip S. Yu. 2008. Truth discovery with multiple conflicting information providers on the web. *IEEE Trans. Knowl. Data Eng.* 20, 6 (2008), 796–808.

[46] Xiaoxin Yin and Wenzhao Tan. 2011. Semi-supervised truth discovery. In *Proceedings of the International World Wide Web Conference (WWW'11)*. 217–226.

[47] Dian Yu, Hongzhao Huang, Taylor Cassidy, Heng Ji, Chi Wang, Shi Zhi, Jiawei Han, Clare Voss, and Malik Magdon-Ismail. 2014. The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding. In *Proceedings of the International Conference on Computational Linguistics (COLING'14)*. 1567–1578.

[48] Hengtong Zhang, Qi Li, Fenglong Ma, Houping Xiao, Yaliang Li, Jing Gao, and Lu Su. 2016. Influence-aware truth discovery. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM'16)*. 851–860.

[49] Bo Zhao and Jiawei Han. 2012. A probabilistic model for estimating real-valued truth from conflicting sources. In *Proceedings of the International Workshop on Quality in DataBases (QDB'12) coheld with VLDB*.

[50] Bo Zhao, Benjamin I. P. Rubinstein, Jim Gemmell, and Jiawei Han. 2012. A Bayesian approach to discovering truth from conflicting sources for data integration. *Proc. VLDB Endow.* 5, 6 (2012), 550–561.

[51] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: Is the problem solved? *Proc. VLDB Endow.* 10, 5 (2017).

[52] Shi Zhi, Fan Yang, Zheyi Zhu, Qi Li, Zhaoran Wang, and Jiawei Han. 2018. Dynamic truth discovery on numerical data. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'18)*. 817–826.

[53] Shi Zhi, Bo Zhao, Wenzhu Tong, Jing Gao, Dian Yu, Heng Ji, and Jiawei Han. 2015. Modeling truth existence in truth discovery. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1543–1552.

Received January 2020; revised July 2020; accepted July 2020