# Empowering Truth Discovery with Multi-Truth Prediction

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

## ABSTRACT

Truth discovery is the problem of detecting true values from the conflicting data provided by multiple sources on the same data items. Since sources' reliability is unknown *a priori*, a truth discovery method usually estimates sources' reliability along with the truth discovery process. A major limitation of existing truth discovery methods is that they commonly assume exactly one true value on each data item and therefore cannot deal with the more general case that a data item may have multiple true values (or *multi-truth*). Since the number of true values may vary from data item to data item, this requires truth discovery methods being able to detect varying numbers of truth values from the multi-source data. In this paper, we propose a multi-truth discovery approach, which addresses the above challenges by providing a generic framework for enhancing existing truth discovery methods. In particular, we redeem the numbers of true values as an important clue for facilitating multi-truth discovery. We present the procedure and components of our approach, and propose three models, namely the *byproduct* model, the *joint* model, and the *synthesis* model to implement our approach. We further propose two extensions to enhance our approach, by leveraging the implications of similar numerical values and values' co-occurrence information in sources' claims to improve the truth discovery accuracy. Experimental studies on real-world datasets demonstrate the effectiveness of our approach.

## Keywords

Truth discovery; multiple truths; empowerment model; value co-occurrence

## 1. INTRODUCTION

Applications in the Big Data era are increasingly relying on the data provided by multiple sources for advanced analytics and decision making. Each day, around 2.5 quintillion bytes of data are generated from various sources, such as sensors in the Internet of Things (IoT) applications, workers in crowdsourcing systems, and transactions in e-Commerce systems [3]. A common requirement of these applications is to handle the multi-source data efficiently and effectively.

Unfortunately, data sources in an open environment are inherently unreliable and the data provided by these sources might be incomplete, out-of-date, or even erroneous [2]. For example, sensors in wild fields may produce inaccurate readings due to hardware limitations or malfunction; weather websites may publish out-of-date weather information due to delayed updates [4]; workers in a crowdsourcing system may assign different labels to the same items as a result of their varying expertise and biases [1]. Moreover, it is not uncommon in e-Commerce systems that sellers provide extremely low prices, which are not actually true, to attract customers [7]. Consequently, given a specific data item, different data sources may provide varying values, which lead to conflicts. This makes it important to detect true values from the conflicting multi-source data to support trusted analytics and reliable decision making.

In general, the multi-source data may contain none, one, or multiple true values on each data item. We recognize the problem of detecting a varying number of true values of each data item from conflicting multi-source data as the *multi-truth discovery problem* (MTD), of which the traditional single-truth discovery problem is a special case. The main challenge regarding MTD is that, given a specific data item, the number of true values is *unknown*. For example, many online stores like textbooksNow.com and textbookx.com list *Miles J. Murdocca* as the only author of the book "*Principles of Computer Architecture*", while other stores, such as A1Books and ActiniaBookstores, post two people, *J Miles Murdocca* and *Heuring P Vincent*, as co-authors of the same book. Given such conflicting records, it is extremely difficult for a user to determine the true authors of the book, as the correct number of authors is also unknown.

Traditional truth discovery methods are unsuitable for MTD as they are designed only for the single-truth scenarios. Given a specific data item, they evaluate a value by assigning it a score. The score is generally a relative measure, with a higher score indicating a higher truth probability of the corresponding value; so when there exists only one true value, they simply take the value with the highest score as the truth. However, when it comes to the case of multiple true values, it becomes impossible to predict the truth with-

*CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA*
© 2016 ACM. ISBN 978-1-4503-4073-1/16/10. . . $15.00
DOI: http://dx.doi.org/10.1145/2983323.2983767

out knowing the exact number of true values (*truth number* for short) as each score itself cannot indicate whether a value is true.

To the best of our knowledge, few studies have focused on the MTD problem. Due to the inherent difficulty of MTD, instead of devising new solutions, we believe it is more feasible to enhance the existing truth methods to cope with the challenge. Based on this insight, we propose an approach that takes into account the number of true values as an important clue to facilitate multi-truth discovery. In a nutshell, we make the following contributions:

- We investigate the characteristics of real-world datasets and thereby propose a multi-truth discovery approach, which takes truth number as an important clue to enhance the existing truth discovery methods. The approach is applicable to various existing truth discovery methods and enables them to deal with MTD.

- We present the procedure and components of the approach and propose three models to implement our approach. The models serve as alternatives each providing a different routine for incorporating the existing truth discovery methods. We further extend the approach by leveraging the co-occurrence information of values in sources' claims and the implication of similar numerical values to improve the truth discovery accuracy.

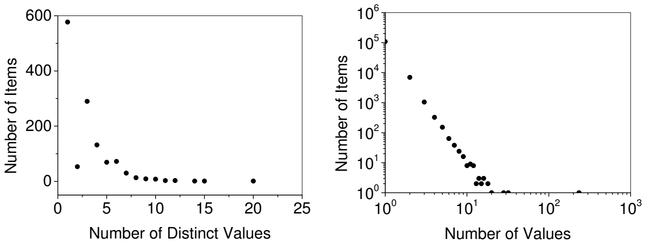
- We conduct experiments on various real-world datasets to evaluate the proposed approach. The results show notable improvement of the existing truth discovery algorithms in accuracy by adopting our approach, without significantly sacrificing the efficiency.

The rest of the paper is organized as follows. We discuss the observations that motivate our work and define the multi-truth discovery problem in Section 2. Section 3 introduces the procedure and basic components of our approach. Section 4 presents the models that implement our approach. Section 5 reports the experimental studies and discussion of the results. We overview the related work in Section 6 and finally, give some concluding remarks in Section 7.

## 2. PRELIMINARIES

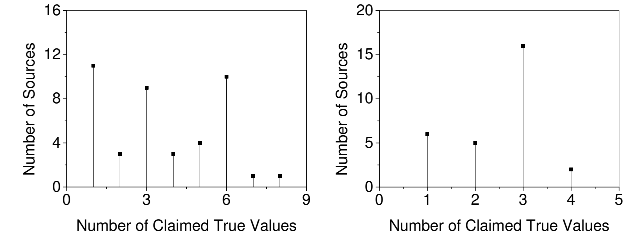
### 2.1 Observations

We investigate the distributions of truth numbers claimed by the sources over data items in various real-world datasets. As an example, Fig. 1a and Fig. 1b show the results on the *author* [17] and *biography* [11] datasets, respectively. In both subfigures, a point $(x, y)$ indicates there exist $y$ data items that have $x$ possible true values in the dataset. The results show a heavy long-tail in the distributions of both datasets, indicating that a significant portion (actually, over a half) of the data items have more than one true value. This means we cannot simply ignore the possible existence of multiple true values on each data item[1] in developing the truth discovery methods. Since the numbers of distinct values claimed by the sources vary among different data items, we cannot presume some fixed truth numbers for all data items either. The only feasible solution is to detect the truth number of each data item dynamically during the truth discovery process.



(a) Author Dataset (b) Biography Dataset

Figure 1: **The number of values claimed on the data items: different data items have varying numbers of values claimed.**



(a) Book (id: 0321334205) (b) Book (id: 0321349946)

Figure 2: **The claimed truth number of sources on two sample data items: the sources are distributed differently over the truth numbers depending on the data items.**

We also investigate the distributions of the claimed truth numbers over data sources regarding different data items of the two datasets. As an example, Fig. 2a and Fig. 2b show the distributions of the claimed truth numbers regarding two books from the author dataset. We observe varying numbers of data sources that claim different numbers of values on the data items. In some cases, the distributions are dominated by some numbers, e.g., more than half of the sources claim three authors for the book (id: 0321349946) (as shown in Fig. 2b); in other cases, no single values dominate the distributions (e.g., in Fig. 2a), which makes the estimation of the correct truth numbers non-trivial.

To further explore the relationship between sources' claims, we investigate the coverage of all claims of sources regarding the same data items of the author dataset. We find that the sources may claim identical, overlapping, or total different sets of values on the same data items. Therefore, in a multi-truth discovery problem, two different claims of sources may not totally support (when they are identical) or oppose (when they are totally different) each other. Instead, they may partially support or oppose each other as their claims may contain some identical values but at the same time some different values. Given such conditions, traditional truth discovery methods would still regard the two claims as mutually exclusive, which leads to inaccurate truth discovery results. For example, it is unreasonable to regard the claims of $s_1$ and $s_7$ (shown in Table 1) as totally exclusive as the values contained in the two claims are almost the same (except $Winston$).

Based on above observations, we obtain the following hints from real-world datasets for effective multi-truth discovery:

[1] We will hereafter use *multi-truth* to denote the *multiple true values on each data item*, for short.

- We should change the truth discovery models in terms of shifting from evaluating the truthfulness[2] of sources' claims to evaluating the truthfulness of individual values. Only in this way the correlation issue between sources' claims could be addressed—that is why we propose the *claim value separation* component in our approach (Section 3.1).
- Sources' claims inherently implicate sources' opinions on the truth numbers of data items. It could be extremely useful to estimate and incorporate the truth numbers into the truth discovery process—that is why we propose the *truth number estimation* component in our approach (Section 3.2).
- Besides truth numbers, the sources' claims contain other implications that cannot be captured by individual values. Such implications could be helpful for achieving more accurate multi-truth discovery, e.g., the co-occurrence of values in the same claims—that is why we propose the *incorporating claim implication* component to extend our approach in Section 4.4.2.
- Since truth numbers belong to numerical values, they have some unique characteristics of the numerical values that could be leveraged to improve the multi-truth discovery—that is why we propose the *incorporating value implication* component to extend our approach in Section 4.4.1.

## 2.2 A Motivating Example

Based on the above insights, we propose a multi-truth discovery approach that takes into account truth numbers to enhance the existing truth discovery algorithms and to enable them to cope with MTD. Now we illustrate the basic ideas of our approach with an example.

EXAMPLE 1. *Suppose we want to collaborate the authors of the book named "Artificial Intelligence: Theory and Practice". Seven websites provide such information but only one of them, $s_1$, provides all true values (Table 1). The problem is challenging since almost every source claims a different set of values and the true and false values are often mixed up in the same claims.*

*Traditional truth discovery methods perform truth discovery at the claim level. For example, the naive voting method would predict {Thomas;Luger} as the truth since this claim is supported by more sources than any other claims—only this claim occurs twice. Such results may not always be reasonable. For example, both James and Yiannis are voted by more sources and should be more likely to be true than Luger.*

*Evaluating each value separately could help alleviate this above issue. This would require reformatting the sources' claims into fine-grained ones each containing only one value, e.g., the claim of $s_1$ would be decomposed into three claims that contain Thomas, James, and Yiannis, respectively (as shown in Table 2a). While improving the truth discovery precision, the reformation, on the other hand, limits the truth discovery methods to discovering only a single value on each data item. For example, using the reformatted claims, the naive voting method can only predict a single value, i.e., Thomas, as the truth.*

Table 1: **A motivating example: seven websites (i.e., sources) provide author information about a book. Only $s_1$ provides all correct authors (i.e., true values).**

| Source | Source's claim | #truth |
|---|---|---|
| $s_1$ | Thomas; James; Yiannis | 3 |
| $s_2$ | Thomas; Luger | 2 |
| $s_3$ | Thomas; James; Winston | 3 |
| $s_4$ | Thomas; Luger | 2 |
| $s_5$ | Thomas; James; Goldberg | 3 |
| $s_6$ | Thomas; Yiannis | 2 |
| $s_7$ | Thomas; James; Yiannis; Winston | 4 |

Table 2: **The sources' claims after reformatting and ranking.**

(a) Reformatted claims

| Source | Source's claim |
|---|---|
| $s_1$ | Thomas |
| $s_1$ | James |
| $s_1$ | Yiannis |
| $s_2$ | Thomas |
| $\cdots$ | $\cdots$ |
| $s_7$ | Winston |

(b) Ranked list

| Value | #vote |
|---|---|
| Thomas | 7 |
| James | 4 |
| Yiannis | 3 |
| Luger | 2 |
| Winston | 2 |
| Goldberg | 1 |

*Our approach enhances existing truth discovery methods by incorporating the consideration of truth numbers into the truth discovery process. Then, it uses these numbers to filter the values and to finally predict the truth. For example, by adopting our approach, the naive voting method estimates the truth number as 3 and ranks the values according to the number of votes they received (Table 2b). It finally predicts the truth as the top-3 values in the ranked list, i.e., {Thomas;James;Yiannis}. Note that there is a tie in the vote counts of 2 and 3. In such cases, the implication of similar numerical values is considered to address the issue. In this example, the truth number is estimated as 3, since 4 is more close to 3 than 2.*

## 2.3 Definitions and Notations

The basic truth discovery problem contains four inputs: *i*) data items, the true values of which are to be discovered, *ii*) data sources, which provide potential true values on data items, *iii*) values, the values claimed by the sources, and *iv*) the associations among the above elements, which indicates which sources claim which values on which data items.

Let $O$ be the set of all data items. For each data item $o \in O$, let $S_o$ be a set of data sources that make claims on $o$ and $V_o$ be the set of distinct values claimed by sources on $o$. The set of all data sources can be represented by $S = \{S_o\}_{o \in O}$. Given a specific value on data item $o$, namely $v_o$, we further denote by $S(v_o)$ the set of data sources that claim $v_o$ on $o$ and $V(s_o)$ the set of values claimed by data source $s_o$ on data item $o$. Generally, there is a *many-to-many* mapping between the elements of $S$ and $\{V_o\}$, meaning each source may claim multiple values on each data item and each value may be claimed by multiple sources.

*Multi-Truth Discovery Problem.* The multi-truth discovery problem distinguishes from the traditional truth discovery problem in that it allows for the detection of varying numbers of true values on the data items. Given a specific dataset, a multi-truth discovery method aims at identifying a set of prospective true values $V_o^e$ from the full set $V_o$ on each

[2] *Truthfulness* takes the value of either true or false.

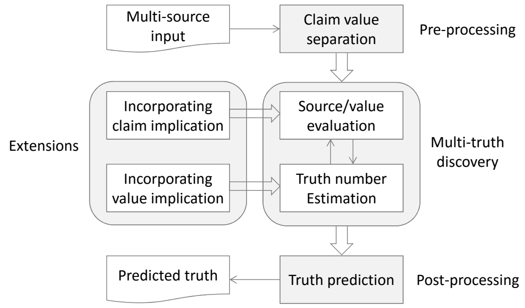


Figure 3: **Multi-truth discovery framework**

data item $o$. $V_o^e$ is expected to be as close to the factual truth (namely $V_o^*$) as possible. A perfect truth discovery would produce a $V_o^e$ that satisfies $V_o^e = V_o^*$.

# 3. APPROACH OVERVIEW

Our approach facilitates multi-truth discovery by incorporating the estimated truth numbers into value evaluation and truth prediction. Fig. 3 shows the procedure and components of our approach, where a broad arrow represents the source supports the destination and a slim arrow indicates information flow. The approach comprises three basic components: *claim value separation*—for pre-processing the multi-source data, *multi-truth discovery*—for evaluating the sources and values along with the estimation of truth numbers, and *truth prediction*—for deriving truth discovery results. Different from the above components, the *extensions* component is the optional part of the approach. For this reason, we will omit the introduction of this component in this section, but present its details in Section 4.4. The following subsections will introduce the three basic components, respectively.

## 3.1 Claim Value Separation

This component prepares the necessary inputs for truth discovery methods. It performs two tasks: *i*) decomposing the original claims of sources into the ones each containing a single value, e.g., reformatting Table 1 into Table 2a, and *ii*) detecting the truth number implicated by each claim on a data item, e.g., the last column of Table 1.

Algorithm 1 formally illustrates the above procedure, which provides the option of incorporating mutual exclusion[3] for truth discovery methods by artificially generating the negative claims (lines 7-8). However, it is worth noting that, unlike the discovery of single-truth, in a multi-truth discovery problem, the mutual exclusion assumption may not be appropriate—given a specific data item, the fact of one value being true does not necessarily exclude the possibility of another value being true. For this reason, incorporating this assumption may not bring more accurate results.

The time complexity of Algorithm 1 is $O(|V||S||O|)$, where $V$ is the maximum number of values claimed on each data item, i.e., $V = \max\{V_o\}_{o\in O}$. Compared with the original claim number, $\sum_{o\in O}|S_o|$, Algorithm 1 increases the number to $\sum_{s\in S_o, o\in O}|V_{s,o}|$.

[3] *Mutual exclusion* is the assumption that the distinct values of the same data item are mutually exclusive.

**Algorithm 1:** Claim Value Separation

```
   Input: the original claims in the multi-source data,
          $\{(s,o,V_{s,o})|V_{s,o} \subseteq V_o \cap V_s, s \in S_o, o \in O\}$
   Output: the set of claims regarding individual values,
           $claimSet$; the claimed truth numbers of the sources on
           each data item, $\#truthSet$.
1  $claimSet \leftarrow \emptyset$
2  $\#truthSet \leftarrow 0$
3  foreach $o \in O$ do
4      foreach $s \in S_o$ do
5          foreach $v \in V_{s,o}$ do
6              $claimSet \leftarrow claimSet \cup \{(s,o,v)\}$
7          foreach $v' \in V_o \backslash V_{s,o}$ do
8              $claimSet \leftarrow claimSet \cup \{(s,o,\neg v')\}$
9          $\#truthSet \leftarrow \#truthSet \cup \{(s,o,|V_{s,o}|)\}$
10 return $claimSet$, $\#truthSet$
```

## 3.2 Multi-Truth Discovery

Traditional truth discovery methods usually perform truth discovery by computing sources' reliability and values' truth probabilities alternately from each other. This component additionally incorporates a *truth number estimation* component for more effective multi-truth discovery. In particular, the truth number is estimated in terms of probabilities. Given a truth number $n$, we denote the truth probability of $n$ by $p(n)$.

Algorithm 2 shows the basic truth discovery procedure, where the algorithm alternately computes the three parameters, $\{\sigma(v)\}$, $\{\tau(s)\}$, and $\{p(n)\}$, from each other. Each parameter can be derived from one or both of the other parameters, although taking more input parameters does not necessarily produce more accurate results. The sequence of the three components, *value evaluation* (lines 3-5), *source evaluation* (lines 6-7), and *truth number estimation* (lines 8-9), may differ, depending on the adopted truth discovery methods.

**Algorithm 2:** Multi-Truth Discovery

```
   Input: the set of reformatted claims $claimSet$; the claimed
          truth numbers of sources $\#truthSet$.
   Output: the evaluation result of each value of each data item;
           the truth probability of each possible truth number of
           each data item.
1  Initialize sources' reliability $\{\tau(s)\}_{s\in S}$
2  do
3      foreach $o \in O$ do
4          foreach $v \in V_o$ do
5              $\sigma(v) \leftarrow$ evaluate $v$
6      foreach $o \in O$ do
7          $\{p_o(n)|n = 1, 2, \cdots, |V_o|\} \leftarrow$ estimate the probability of
           every possible truth number of $o$
8      foreach $s \in S$ do
9          $\tau(s) \leftarrow$ evaluate $s$
10 while non-convergence;
11 return $\{\tau(s)\}_{s\in S}$, $\{\sigma(v)\}_{v\in V_o, o\in O}$, $\{p_o(n)\}_{n=1,2,\cdots,|V_o|}$
```

Note that, in this section, we simply describe the basic procedure. The details about how the probabilities of truth numbers are incorporated into source/value evaluation are introduced along with the different models in Section 4. The complexity of Algorithm 2 is $O(|M||V||S||O|)$, where $M$ is the number of iteration. This complexity may increase in order, depends on the methods used for calculating the three parameters.

## 3.3 Truth Prediction

While different implementations of our approach may differ in their truth discovery methods, their final steps are common, i.e., *truth prediction*. This component predicts and outputs true values of each data item as the final results. The prediction is based on ranking the values of each data item according to their evaluation results. Intuitively, the values with higher evaluation scores should have higher truth probabilities; so given a truth number $n$, the top-$n$ values in the ranked list should be predicted as the truth.

Algorithm 3 illustrates the truth prediction procedure, where the true values are predicted based on the truth probabilities of truth numbers estimated by Algorithm 2. In particular, given a data item $o$ and a value ranked at the $k$-th place in the list—meaning it is the $k$-th most promising value—its truth probability is calculated as $\sum_{i=k}^{|V_o|} p_o(i)$. The value is regarded as true if the predicted truth number is larger than or equal to $k$. In this way, the ranking methods enable truth discovery methods to detect varying numbers of true values regardless the values' evaluation results are relative measures of truthfulness or not. This feature relieves the truth discovery methods from the necessity of incorporating mutual exclusion to enable multi-truth discovery. The time complexity of Algorithm 3 is $O(|V_o| \log |V_o|)$.

```
Algorithm 3: Truth Prediction
Input: the evaluation result of each distinct value on each data
       item of the multi-source data, $\{\sigma(v_o)|v_o \in V_o, o \in O\}$;
       the truth probabilities of every possible truth number of
       every data item, $\{p_o(n)|n = 1, 2, \cdots, |V_o|\}$.
Output: the set of predicted true values of each data item.
foreach $o \in O$ do
    Sort $\{v_o\}_{v_o \in V_o}$ in descendant order by $\sigma(v_o)$
    // Suppose the sorting result is $\{v_o^1, v_o^2, \cdots, v_o^{|V_o|}\}$
    $V_o^e \leftarrow \emptyset$
    $sum \leftarrow 1$
    foreach $n = 1, 2, \cdots, |V_o|$ do
        if $sum > 0.5$ then
            $V_o^e \leftarrow V_o^e \cup \{v_o^n\}$
        $sum \leftarrow sum - p_o(n)$
return $\{V_o^e|o \in O\}$
```

# 4. TRUTH DISCOVERY MODELS

Based on the discussion in Section 3.2, there can be various implementations of the multi-truth discovery methods for our approach. In this section, we introduce three simple yet effective models that incorporate truth number as a new clue for multi-truth discovery. In particular, the byproduct model is designed to incur the minimal interference with the existing truth discovery methods; so it is the easiest to use and requires little efforts to apply. The second requires a little more while the third requires the most modifications to the original truth discovery methods

## 4.1 Byproduct Model

The straightforward way of incorporating truth numbers is to separate truth number prediction from source/value evaluation. This model does not interfere with the truth discovery methods but only adds an additional step to predict truth numbers based on sources' evaluation results.

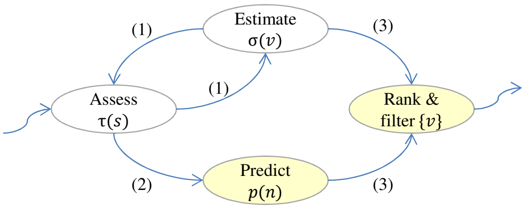


Figure 4: **Truth number as byproduct**

Fig. 4 illustrates the basic ideas of this model, where each node represents a functional component and the arrows denote the information flow. Specially, the arrow with no source denotes the initial inputs of the truth discovery method (e.g., in Fig. 4, the initial inputs are the prior probabilities of sources) and the arrow with no destination denotes the final outputs of the truth discovery method. The shaded nodes are only performed once, while the other nodes belong to the truth discovery method and may participate the iteration. Specially, the exit node (i.e., *rank & filter*) corresponds to the truth prediction method (Algorithm 3), while all the other nodes belong to the multi-truth discovery method (Section 3.2).

The byproduct model involves three steps, as indicated by the numbers beside the arrows in Fig. 4. We describe these steps as follows:

1. The sources and values are evaluated alternately through an iterative procedure, yielding $\{\tau(s)\}$ and $\{\sigma(v)\}$.

2. $\{p(n)\}$ are predicted based on sources' reliability. Suppose $n_{s,o}$ is the truth number claimed by source $s_o$ ($s_o \in S_o$) on data item $o$. The unnormalized truth probability of $p_o^*(n)$ is estimated by:

$$p_o^*(n) = \sqrt[|S_o|]{\prod_{n_{s,o}=n} \tau(s) \prod_{n_{s,o} \neq n} (1 - \tau(s_o))} \quad (1)$$

where $n \in \{1, 2, \cdots, |V_o|\}$. The resulting values are then normalized to represent probabilities:

$$p_o(n) = \frac{p_o^*(n)}{\sum_{i=1}^{|V_o|} p_o^*(i)} \quad (2)$$

3. Algorithm 3 is applied to produce the truth discovery results.

## 4.2 Joint Model

This model differs from the byproduct model in that the truth number estimation is incorporated into the source/value evaluation process (Fig. 5). During the iteration, this model uses the same methods as the byproduct model (i.e., Eq. (1) and Eq. (2)) to estimate the probabilities of truth numbers based on sources' reliability. But it additionally uses the estimated probabilities of truth numbers to reevaluate the values by applying Algorithm 3.

Note that, we would not use the truth probabilities of values to infer the probabilities of truth numbers for two reasons. First, different from sources' reliability, the evaluation results of values are not always in the form of probabilities, depending on the truth discovery methods adopted. Second, such inference incurs severe computational overhead, which makes the approach unfeasible. For example, given truth probabilities of values on data item $o$, say $\{\sigma(v_o)\}$, the truth probability of a truth number $n$, $p_o(n)$, is calculated by:

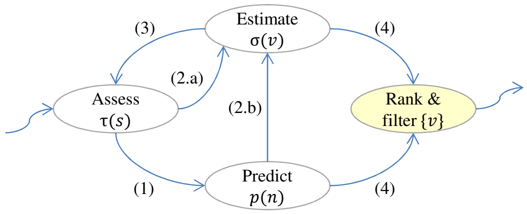


Figure 5: **Jointly inferring truth probability**

$$p_o(n) = \sum_{V_o^{(n)} \in C_o^{(n)}} \prod_{v \in V_o^{(n)}} \sigma(v) \prod_{v' \in V_o \setminus V_o^{(n)}} (1 - \sigma(v')) \quad (3)$$

where $C_o^{(n)}$ is the set of all different combinations formed by $n$ distinct values of $V_o$; $V_o^{(n)}$ is one of these combinations, satisfying $|V_o^{(n)}| = n$ and $V_o^{(n)} \subseteq V_o$. Based on Eq. (3), we can estimate the time complexity of inferring $\{p_o(n)\}$ as:

$$O(\sum_{i=1}^{|V_o|} C(|V_o|, i) \cdot |V_o|) = O(2^{|V_o|}|V_o|) \quad (4)$$

where $C(|V_o|, i)$ is the number of $i$-combinations formed from $|V_o|$ distinct values.

## 4.3 Synthesis Model

This model (Fig. 6) separately evaluates sources' reliability based on two different types of inputs, i.e., the evaluation results of values and the estimated truth probabilities of truth numbers. The evaluation results of sources are obtained by synthesizing the results of the above separate evaluations. Besides using Eq. (1) and Eq. (2) in Step 1, and Algorithm 3 in Step 4, this model additional requires: *i*) assessing sources' reliability based on the probabilities of truth numbers (Step 2.b), and *ii*) synthesizing the separate evaluation results of sources' reliability (Step 3).

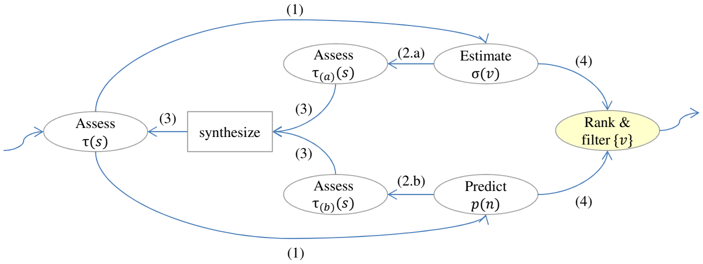


Figure 6: **Synthesizing evaluations of source**

We present the methods that perform the above calculations as follows:

- Given a source $s$ and the set of related data items (on which $s$ has claimed values), namely $O(s)$, suppose $V_o(s)$ is the set of values claimed by $s$ on a data item $o$. The reliability of $s$ is estimated from $\{p(n)\}$ by:

$$\tau_{(b)}(s) = \frac{1}{|O(s)|} \sum_{o \in O(s)} p(|V_o(s)|) \quad (5)$$

- The separate evaluation results of a source $s$ can be synthesized to a single result by taking the weighted sum:

$$\tau(s) = \lambda \cdot \tau_{(a)}(s) + (1 - \lambda) \cdot \tau_{(b)}(s) \quad (6)$$

where the $\lambda$ value can be tuned to achieve the optimal results. By default, we have no bias towards the two aspects and set $\lambda = 0.5$.

## 4.4 Extensions

In this section, we present two methods that can be incorporated into the multi-truth discovery models to improve the accuracy.

### 4.4.1 Incorporating Value Implication

Since truth numbers are numerical values, we could leverage the unique characteristics of numerical values to improve value evaluation. In particular, numerical values are unique in that different values may have an influence on the truthfulness of each other. It is common sense that similar numerical values are more likely to have similar truth probabilities. For example, a higher truth probability of the truth number 6 could greatly increase the truth probability of the similar values like 5 and 7, but would have only a minor or even negative influence on the truth probability of a distance value like 1.

Incorporating such implication of values can generally help improve the robustness of truth discovery algorithms in terms of smoothing the fluctuations and anomalies in the evaluation results[4]. Intuitively, the truth discovery results cannot be justified if two very similar values are assigned significantly different truth probabilities.

Given a truth number $n_o$ of data item $o$, we use a similar method as those used in previous works [17, 5] to amend the estimated truth probability of a truth number $n_o$:

$$p^*(n_o) = p(n_o) + \rho_n \cdot \sum_{n'_o \in \{1,2,\cdots,|V_o|\} \setminus \{n_o\}} sim(n'_o, n_o) \cdot p(n'_o) \quad (7)$$

where $sim(\cdot) \in [0, 1]$ measures the similarity between two values, $p(n_o)$ and $p^*(n_o)$ are the estimated truth probabilities of truth number $n_o$ before and after the amendment, respectively, and $\rho_n \in (0, 1]$ represents the influence strength. The amending results should be normalized to ensure that the truth probabilities of all possible truth numbers of each data item sum up to 1. The time complexity of this component is $O(\sum_{o \in O} |V_o|^2) \leq O(|O||V|^2)$, where $V = \max_{o \in O} |V_o|$.

### 4.4.2 Incorporating Claim Implication

Besides the truth numbers, another type of information is unpreserved during the claim reformation—the *co-occurrence* of values in the same claims. Intuitively, the values in the same claims are likely to have similar truth probabilities. Similar to the implication between numerical values, the values that co-occur in the same claims may also have an influence each other.

In this section, we aim at quantifying this influence for more accurate value evaluation. In particular, for each pair of distinct values on a data item, we find out the sources that contain both values in the same claims. For example, such source for the example in Table 1 is shown as a matrix in Table 3. The matrix is equal to a graph with values

[4] The fluctuating truth probability distributions of existing truth discovery algorithms over the distinct values for the same items are investigated by [15].

as vertices and sets of sources as weights. We initialize the weight of each edge as the sum of the reliability of the corresponding sources and normalize the weights to ensure every column sums to 1. We also add edges with small weights between the unconnected values to ensure full connectivity and use page-rank algorithms to reach a stationary state over the graphs. The adjacent matrix of the graph should be stochastic, irreducible, and aperiodic, and is guaranteed to converge [13].

Based on the above results, we incorporate the influence between co-occurring values in a similar way as we incorporate the value implication in Section 4.4.1. Given a value $v_o$ on data item $o$, its evaluation result is amended by:

$$\sigma^*(v_o) = \sigma(v_o) + \rho_c \cdot \sum_{i=1}^{|V_o|} w(v_o, v_o^{(i)})\sigma(v_o^{(i)}) \tag{8}$$

where $w(\cdot) \in [0,1]$ is the weight between two values, and $\rho_c \in (0,1]$ represents the influence strength. The time complexity of this component is also $O(|O||V|^2)$.

# 5. EXPERIMENTS

In this section, we report the experimental studies on the comparison of our approach with the state-of-the-art algorithms and the impact of the extensions on the performance of our approach, using three real-world datasets.

## 5.1 Experimental Setup

### 5.1.1 The Datasets

We employed three real-world datasets in our experiments:

- The author dataset [17] contains 33,971 records crawled from www.abebooks.com. Each record represents a claim of a bookstore on the authors of a book. We removed the invalid and duplicated records, as well as the records with only minor conflicts to make the problem more challenging. We finally obtained 12,623 distinct claims describing 649 sources (i.e., bookstores) claiming authors for 664 books. On average, each book has 3.2 authors. The ground truth provided for the original dataset is used as gold standard.

- The biography dataset [11] contains 11,099,730 editing records about people's birth/death dates, spouses, and parents/children on Wikipedia. We specially focus on people's children information and obtained records about 55,259 users claiming children for 2,579 people. In the resulting dataset, each person has on average 2.45 children. For experimental purposes, we used the latest editing records as the ground truth.

- The director dataset was prepared by ourselves via crawling 33,194 records from 16 major movie websites. The records describe the associate between 1,081 directors and 6,402 movies. On average, each director is associated with 8.8 movies in the dataset. We sampled 100 directors and extracted their movie lists from Wikipedia as the ground truth.

### 5.1.2 Evaluation Metrics

We evaluated the performance of the algorithms using four measures: three for measuring accuracy, and one for efficiency:

- *Precision* and *recall*. In multi-truth discovery, the precision/recall regarding each data item falls in the ranges of $(0,1)$. Given $M$ runs of an algorithm, its precision and recall are calculated by:

$$\begin{cases} precision = \frac{1}{M|O|}\sum_{m=1}^{M}\sum_{o\in O}\frac{V_o^* \cap V_o^{(m)}}{V_o^{(m)}} \\ recall = \frac{1}{M|O|}\sum_{m=1}^{M}\sum_{o\in O}\frac{V_o^* \cap V_o^{(m)}}{V_o^*} \end{cases} \tag{9}$$

  where $\mathcal{V}_o^{(m)}$ is the set of true values predicted by the $m$-th run of the algorithm for data item $o$.

- $F_1$ *score*. We evaluated the overall accuracy of truth discovery by the harmonious mean of precision and recall, i.e., $F_1$ *score*. It is calculated by:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{10}$$

- *Execution time*. Beside accuracy, we used execution time to measure the impact of our approach on the efficiency of truth discovery algorithms:

$$time = \frac{1}{M}\sum_{m=1}^{M} time^{(m)} \tag{11}$$

### 5.1.3 Baseline Methods

We classified existing truth discovery methods into five categories and selected the typical and advantageous ones from each category to evaluate our approach:

- **Primitive method**, i.e., *Voting*. For each item, it outputs the values contained by the most frequently occurring claim as the estimated truth without iteration.

- **Iterative methods**, i.e., *Avg-Log* [11], *2-Estimates* [6], *TruthFinder* [17], and *Accu* [5]. They all evaluate sources and values alternately from each other but use different calculation methods. Specially, *2-Estimates* considers both sources' claims and their negative implications (by incorporating mutual exclusion).

- **Optimization method**, i.e., the *Conflict Resolution on Heterogeneous Data (CRH)* framework [9], which models truth discovery as the problem of minimizing the weighted deviation of multi-source inputs from the estimated truth.

- **Probabilistic method**, i.e., the *Latent Truth Model (LTM)* [19], which applies generative models to estimate truths. It differs from other existing methods in taking the reformatted instead of the original claims as input.

- **Multi-truth discovery method**, i.e., the *Multi-truth Bayesian Model (MBM)* [16], which comprehensively incorporates a new mutual exclusion definition for multi-truth discovery from the reformatted claims.

We directly used the source code of CRH in Matlab and implemented all other algorithms using Java SDK 8. We conducted a series of test-runs to determine the optimal parameter settings for the baseline methods. All experiments were conducted on a 64-bit Windows 10 PC with an octa-core 3.4GHz CPU and 32GB RAM.

Table 3: **The sources that make claims about the co-occurrence of different values.**

| | D. Thomas | A. James | A. Yiannis | G. Luger | P. Winston | D. Goldberg |
|---|---|---|---|---|---|---|
| D. Thomas | – | $s_1, s_3, s_5, s_7$ | $s_1, s_6, s_7$ | $s_2, s_4$ | $s_3, s_7$ | $s_5$ |
| A. James | $s_1, s_3, s_5, s_7$ | – | $s_1, s_7$ | – | $s_3, s_7$ | $s_5$ |
| A. Yiannis | $s_1, s_6, s_7$ | $s_1, s_7$ | – | – | $s_7$ | – |
| G. Luger | $s_2, s_4$ | – | – | – | – | – |
| P. Winston | $s_3, s_7$ | $s_3, s_7$ | $s_7$ | – | – | – |
| D. Goldberg | $s_5$ | $s_5$ | – | – | – | – |

## 5.2 Comparative Studies

Table 4 shows the performance of different algorithms before and after adopting our approach in terms of precision, recall, and the $F_1$ score on the three datasets. In this part of experiments, we only show the results achieved by the joint model of our approach (Section 4.2) for the comparison, but leave the comparison of different models of our approach to the next section. In particular, each cell of Table 4 contains two values, i.e., the performance values achieved before and after adopting our approach, respectively. We can see that our approach not only enables these algorithms to detect multiple true values from the reformatted claims but also significantly improves these algorithms in both precision and recall. The average percentages of improvement on different metrics and datasets by adopting our approach range from 3% to 20%. This indicates that our approach is effective in improving the accuracy of multi-truth discovery using existing truth discovery algorithms on the experimental datasets.

Fig 7a shows the performance of the algorithms before and after adopting our approach in terms of execution time on the movie datasets (see the columns with strips). We omit to show the results on other two datasets as they all lead to similar conclusions. The results show that adopting our approach (any one of the three implementations) does not significantly decrease the truth discovery efficiency. This can be attributed by two reasons. First, the basic components of our approach inherently not computation intensive, as manifested by their time complexity analysis. Therefore, they incur only neglectable computation overhead when compared with the computation amount of truth discovery algorithms themselves. Second, the number of distinct values in the real-world datasets are relatively small; so the extension components also incur only minor computation overhead.

## 5.3 Impact of Different Concerns

In this section, we report the experiments to compare the different implementations of our approach, as well as the studies on the impact of the extension components on the performance of our approach.

### *5.3.1 Impact of Implementation Models*

Figure 7a shows the performance of different implementations of our approach in terms of execution time. The results show no significant difference between the execution time of the three models of our approach. The reason is that the execution time of all the three algorithms is determined by the performance of the extension components, whose complexity is, in turn, determined by the problem scale. Therefore, given the same truth discovery problem, the different models only slightly differ in their execution time.

Figure 7b shows the performance of the three implementations of our approach (incorporated with the Avg-Log algorithm) in terms of precision and recall on the three datasets. The results show that, although all the three models improve the truth discovery accuracy, the effect of the byproduct model is limited. Since truth numbers are derived from and only used after obtaining the source/value evaluation results, they merely serve as a post-processing step and do not affect the evaluation results. The superior accuracy of the other two models suggests that incorporating truth numbers closely to the truth discovery process can generally lead to better evaluation results and further to better truth discovery results. But, on the other hand, the byproduct model has the advantage of incurring only the minimal changes/extensions to the truth discovery algorithms.

While the joint model changes the method for value evaluation by using ranked valued and probabilities of truth numbers, the synthesis model, in contrast, changes how the sources are assessed via considering the correctness of their claimed truth numbers. In fact, both models explicitly (the joint model) or implicitly (the synthesis model) take into account the correctness of source-claimed truth numbers in evaluating the sources. This is especially important for multi-truth discovery: without such consideration, most truth discovery methods tend to favor the sources that claim fewer values by assigning them with higher evaluation results. Also, it conforms to our intuition that the sources claiming the correct numbers of truth values generally have a higher chance of providing the true values. On the other hand, since the traditional source quality refers only to sources' precision in their provided values, combining the traditional source quality with sources' precision in their claimed truth numbers in a hard way (like it is performed in the synthesis model) may not always be appropriate. For this reason, though achieving similar performance, the joint model seems more stable and insensitive to the specific datasets than the synthesis model.

### *5.3.2 Impact of Extensions*

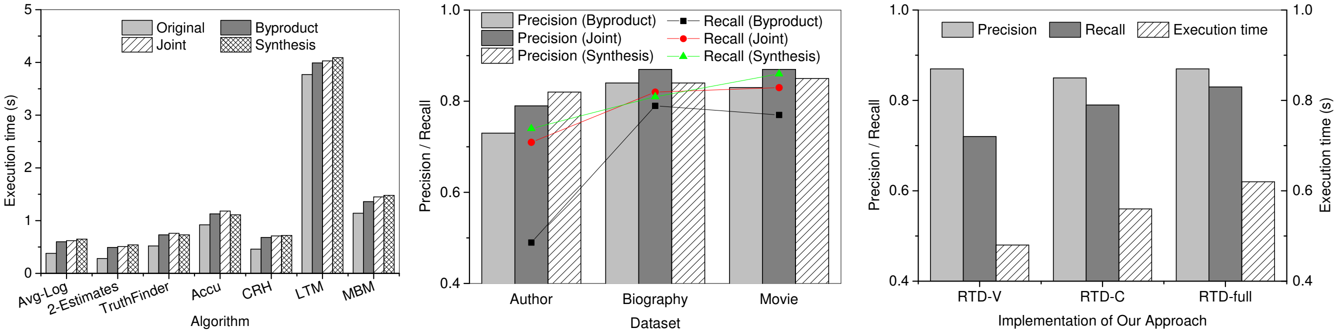
To evaluate the impact of the two extension components, we derive two variants of our approach and name the joint model that incorporates both extensions *RTD*, for ease of illustration:

- *RTD-V*: a version of *RTD* that only incorporates value implication (Section 4.4.1) to extend the basic model.
- *RTD-C*: a version of *RTD* that only incorporates claim implication (Section 4.4.2) to extend the basic model.

Fig. 7c shows the performance of the above three implementations of our approach (incorporated with the Avg-Log algorithm) on the movie dataset. We omit to show the results on other datasets since shifting to another model or dataset would not affect the conclusions drawn. The results show a positive effect of leveraging the implications of similar numerical values for calibrating the estimated truth numbers. This conclusion is consistent with those drawn by

Table 4: **Accuracy comparison of different algorithms before and after adopting our approach (joint model) on three real-world datasets. The best performance values on each dataset are bolded.**

| Method | Author dataset | | | | | | Biography dataset | | | | | | Director dataset | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Precision | | Recall | | $F_1$ score | | Precision | | Recall | | $F_1$ score | | Precision | | Recall | | $F_1$ score | |
| Voting | **0.88** | 0.86 | 0.49 | 0.62 | 0.63 | 0.72 | 0.87 | 0.84 | 0.52 | 0.63 | 0.65 | 0.72 | 0.85 | 0.83 | 0.64 | 0.77 | 0.74 | 0.80 |
| Avg-Log | 0.70 | 0.79 | 0.38 | 0.71 | 0.49 | 0.75 | 0.84 | 0.87 | 0.79 | 0.82 | 0.81 | 0.84 | 0.85 | 0.87 | 0.69 | 0.83 | 0.78 | 0.85 |
| 2-Estimates | 0.79 | 0.86 | 0.65 | 0.80 | 0.71 | 0.83 | 0.84 | 0.84 | 0.62 | 0.80 | 0.71 | 0.82 | 0.85 | 0.85 | 0.77 | 0.86 | 0.83 | 0.85 |
| TruthFinder | 0.73 | 0.86 | 0.80 | 0.83 | 0.76 | 0.84 | 0.81 | 0.87 | **0.85** | 0.81 | 0.83 | 0.84 | 0.83 | 0.85 | 0.86 | 0.86 | 0.86 | 0.85 |
| Accu | 0.79 | 0.82 | 0.65 | 0.80 | 0.71 | 0.81 | 0.84 | **0.89** | 0.62 | 0.79 | 0.71 | 0.84 | 0.83 | 0.87 | 0.69 | 0.86 | 0.75 | 0.86 |
| CRH | 0.73 | 0.86 | 0.65 | 0.83 | 0.69 | 0.84 | 0.81 | 0.84 | 0.58 | 0.79 | 0.67 | 0.81 | 0.85 | **0.89** | 0.72 | **0.88** | 0.78 | **0.88** |
| LTM | 0.84 | **0.88** | 0.75 | 0.83 | 0.79 | **0.85** | 0.87 | **0.89** | 0.82 | **0.85** | 0.84 | **0.87** | 0.83 | **0.89** | 0.79 | 0.83 | 0.83 | 0.86 |
| MBM | 0.79 | 0.84 | 0.83 | **0.85** | 0.80 | 0.84 | 0.84 | 0.89 | **0.85** | **0.85** | 0.84 | **0.87** | 0.81 | **0.89** | 0.86 | **0.88** | 0.85 | **0.88** |



(a) Efficiency

(b) Accuracy

(c) Extensions

Figure 7: **Comparison of different algorithms in terms of precision, recall, and execution time on real-world datasets.**

previous works [17, 5]. The effect of incorporating the implication of values' co-occurrence in sources' claims is also evident. It conforms to our intuition that the chance is much higher for two true values to co-occur than the chance that two false values or one true value and one false value to co-occur in the same claim. Given most sources in the real-world are in most cases reliable, the chance is even lower for two or more false values to co-occur in the same claim. In fact, this extension interprets the correlation among the co-occurring values in a more general way, which takes the interpretation of the traditional truth discovery methods and the interpretation of the methods based on the reformatted claims as two extreme cases. Instead of forcing all the values in the same claim to have the same truth probability (like what the traditional methods do) or evaluating those values totally independently (like what the methods based on the reformatted claims do), our approach quantifies the mutual influence of such values for more accurate truth discovery.

# 6. RELATED WORK

As a major challenge of integrating the Big Data, truth discovery has been actively researched in recent years, and fruitful results have been delivered [10].

Traditional truth discovery methods typically include the iterative methods [17, 11, 6, 5], maximum likelihood estimation (MLE) methods [14], optimization methods [18, 9, 8], and Bayesian methods [19, 12]. In particular, the iterative methods alternately evaluate sources and values through an iterative procedure. The MLE methods perform evaluations by maximizing the likelihood of observations on sources' claims. The optimization methods define and solve the truth discovery problem as an optimization problem. Finally, the Bayesian methods define probabilistic models, esp. probabilistic graphic models, which perform maximum a posterior probability (MAP) optimization for truth discovery. A common issue with the traditional methods is that they assume exactly one true value for each data item. This assumption, however, does not always stand in real-world scenarios as many data items have multiple true values (or *multi-truth*). In fact, the traditional problem of discovering the single true value is just a special case of the multi-truth discovery problem. The existing methods are unsuitable for multi-truth discovery as they commonly perform truth discovery at the claim level and therefore cannot well handle the correlation among sources' claims. This could significantly degrade the truth discovery accuracy.

Unfortunately, until now, there are very few works on the multi-truth discovery. Zhao *et al.* [19] first propose to decompose sources' claims into fine-grained ones each regarding a single value, so that multi-truth could be identified by determining the truthfulness of each value individually. Wang *et al.* [16] also use a binary probabilistic model that is capable of detecting multi-truth. Though doable, the binary probabilistic model has several disadvantages. First, it is not a generic approach, meaning it only fits for the probabilistic methods and no other probabilistic models except the binary model (represented by the Bernoulli distribution) is accepted. This limits the number and diversity of the methods applicable for multi-truth discovery. Second, it requires incorporating mutual exclusion (by adding negative claims to the original claim set) to obtain absolute evaluation results of values, so as to enable the determination of each value's truthfulness individually. However, as aforementioned (see Section 3.1), the mutual exclusion assumption may not always be appropriate in a multi-truth discovery problem as the sources may only provide partial truth. In addition, incorporating mutual exclusion makes the approach sensitive to the distribution of the positive/negative claims. If the case that most sources claim incomplete truth, which is very common in the real-world, the negative claims could overwhelm the positive claims in number, leading to the low recall of truth discovery methods.

In comparison, our approach has the following advantages. First, it is a generic framework that is applicable to any truth discovery method, as long as it can produce evaluation results of sources and values, regardless the evaluation results are absolute measures (i.e., in the form of probabilities) or relative measures. All the four categories of traditional truth discovery methods, as well as those incorporating mutual exclusion, can be incorporated into our approach. Second, as demonstrated by our experiments, our approach only slightly affects the efficiency of the original truth discovery methods but generally achieves significantly higher accuracy. Third, our approach is flexible to incorporate more clues to further improve the truth discovery accuracy. For example, the unique characteristics of numerical values are leveraged by our approach to enforce more consistent truth number estimation.

## 7. CONCLUSION

In this paper, we have conducted a focused study on the problem of detecting varying numbers of truth values for each data item, or the multi-truth discovery problem (MTD). Although there widely exist data items that have multiple true values, MTD is rarely studied by previous efforts. We have proposed an approach for empowering the existing truth discovery methods to address the problem. In particular, we presented the basic procedure and functions that enable the existing truth discovery methods to handle the MTD problem. Three models and their corresponding algorithms are provided as the alternative routines for incorporating existing truth discovery methods in our approach. Truth numbers are predicted and leveraged as an important clue for improving the truth discovery accuracy, with the implications of numerical values and the co-occurring values in claims leveraged as the extension components to further enhance our models. The approach is applicable to all known existing truth discovery methods with low computational overhead. Experiments on three real-world datasets demonstrate the effectiveness of the approach.

## 8. REFERENCES

[1] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, (2):76–81, 2013.

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

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

[4] X. L. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava. Global detection of complex copying relationships between sources. *Proc. the VLDB Endowment*, 3(1-2):1358–1369, 2010.

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

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

[7] D. J. Kim, D. L. Ferrin, and H. R. Rao. A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. *Decis. Support Syst.*, 44(2):544–564, 2008.

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

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

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

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

[12] J. Pasternack and D. Roth. Latent credibility analysis. In *Proc. the 22th international conference on World Wide Web (WWW)*, pages 1009–1020, 2013.

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

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

[15] X. Wang, Q. Z. Sheng, X. S. Fang, X. Li, X. Xu, and L. Yao. Approximate truth discovery via problem scale reduction. In *Proc. the 24th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 503–512, 2015.

[16] X. Wang, Q. Z. Sheng, X. S. Fang, L. Yao, X. Xu, and X. Li. An integrated bayesian approach for effective multi-truth discovery. In *Proc. the 24th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 493–502, 2015.

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

[18] X. Yin and W. Tan. Semi-supervised truth discovery. In *Proc. the 20th international conference on World Wide Web (WWW)*, pages 217–226, 2011.

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