An Integrated Bayesian Approach for Effective Multi-Truth Discovery

Xianzhi Wang,
Quan Z. Sheng,
Xiu Susie Fang, Lina Yao
School of Computer Science
The University of Adelaide
Adelaide, SA 5005, Australia
xianzhi.wang@adelaide.edu.au

Xiaofei Xu
School of Computer Science
and Technology
Harbin Institute of Technology
Harbin 150001, China
xiaofei@hit.edu.cn

Xue Li
School of IT and Electrical
Engineering
The University of Queensland
Brisbane, QLD 4072, Australia
xueli@itee.uq.edu.au

ABSTRACT

Truth-finding is the fundamental technique for corroborating reports from multiple sources in both data integration and collective intelligent applications. Traditional truth-finding methods assume a single true value for each data item and therefore cannot deal will multiple true values (i.e., the *multi-truth-finding* problem). So far, the existing approaches handle the multi-truth-finding problem in the same way as the single-truth-finding problems. Unfortunately, the multi-truth-finding problem has its unique features, such as the involvement of sets of values in claims, different implications of inter-value mutual exclusion, and larger source profiles. Considering these features could provide new opportunities for obtaining more accurate truth-finding results. Based on this insight, we propose an integrated Bayesian approach to the multi-truth-finding problem, by taking these features into account. To improve the truth-finding efficiency, we reformulate the multi-truth-finding problem model based on the mappings between sources and (sets of) values. New mutual exclusive relations are defined to reflect the possible co-existence of multiple true values. A finer-grained copy detection method is also proposed to deal with sources with large profiles. The experimental results on three real-world datasets show the effectiveness of our approach.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Management—*Data Mining*; I.2.m [**Computing Methodologies**]: Artificial Intelligence—*Miscellaneous*

General Terms

Models; Algorithms; Experimentation; Measurement

CIKM'15, October 19–23, 2015, Melbourne, Australia.
© 2015 ACM. ISBN 978-1-4503-3794-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2806416.2806443.

Keywords

Truth discovery; multi-truth-finding features; Bayesian model; data source dependence

1. INTRODUCTION

Integrating data from multiple sources has been increasingly becoming a commonplace in both Web and the emerging Internet of Things (IoT) applications to support collective intelligence and collaborative decision making [4]. Unfortunately, it is not unusual that the information about a single item comes from different sources, which might be noisy, out-of-date, or even erroneous. It is therefore of paramount importance to resolve such conflicts among the data and to find out which piece of information is more reliable [15]. For example, in a recent controversy on Obama's birthplace[1], some people rumored Kenya, while others insisted on Hawaii. Clearly, such conflicts can be extremely disturbing and misleading to the users who want to find the specific facts on something or somebody they concern [12]. Solutions to this challenge are generally recognized as *truth-finders*. Different from methods that seek non-factual truth (e.g., aggregating users' rating on a product, or analyzing people's opinions on a recent event), truth-finders aim at discovering the factual truth, such as the birthplace of Obama and the capital city of the United States.

While the *single-truth-finding problem* (STF)—which aims at finding the single true value for an item—has been widely studied, a more general case, where multiple true values (or multi-truth) might exist for a single item, is rarely explored [24]. In fact, multi-truth scenarios commonly exist in our real lives. For example, a book is usually authored by several people; a conference may have several deadlines; and the presidents of the United States involve a long list of names. We recognize the discovery of multiple true values (for either one or multiple data items) as the *multi-truth-finding problem* (MTF), of which STF can be treated as a special case. We identify the main challenges on solving MTF as follows:

- *Unknown quality on data sources.* The quality of data sources (e.g., trustworthiness) varies and is usually unknown *a priori* to truth-finding methods. For example, no website guarantees how much information it

[1] http://beforeitsnews.com/obama-birthplace-controversy/

publishes is accurate. Without assessing and differentiating the quality of data sources, truth-finding approaches could be easily misled by low quality and dependent sources[2].

- *Poor availability of the ground truth.* Obtaining the ground truth is a nontrivial task by itself due to the possibly numerous sources and data items involved. For example, it took tremendous efforts to setup a gold standard for the experimental book-author dataset—by manually checking the covers of each book [21], not to mention the more practical tasks like consolidating millions of book records from different libraries. The difficulty in obtaining the ground truth suggests an *unsupervised* approach to the truth-finding problem.

- *Unique multi-truth-finding features.* The multi-truth-finding problem (MTF) has unique features that should be properly addressed. For instance, instead of being totally different or exactly the same, the values claimed by different sources may overlap. Also, claiming one value for an item does not necessarily imply disclaiming all the other values for the item, because the claimed value may only cover the truth partially. All these features require special consideration when developing truth-finding solutions.

In this paper, we propose an integrated Bayesian approach to address the above challenges. In a nutshell, we make the following main contributions:

- We propose to reformulate the problem model for multi-truth discovery based on the relations between sources and values, and present corresponding methods for grouping sources and values to enable the reformulation. The reformulation can significantly reduce the computation load when solving the multi-truth-finding problem, without sacrificing the accuracy of the truth-finding results.

- We develop an integrated Bayesian model, which comprehensively incorporates novel methods on three key aspects, namely *source/value grouping*, *source dependency*, and *inter-value mutual exclusion*, to solve MTF. In particular, we define a method for calculating the effect of mutual exclusion between different values, by taking into account the agreement occurring by chance, similar to what *Kappa coefficient* does [6]. We also develop a finer-grained copy detection method to infer source dependencies. The new method is more efficient and especially suitable for sources with large profiles (i.e., sources that claim lots of values).

- We empirically show that our approach outperforms traditional methods using three large real-world datasets. We also study the impact on the effectiveness of the proposed approach of the three technical aspects in the Bayesian model.

The rest of the paper is structured as follows. Section 2 reviews the related work. Section 3 reformulates the multi-truth-finding problem. Section 4 presents the details of our solution, including the integrated Bayesian model and the related algorithms. Section 5 reports our experimental results. Finally, Section 6 provides some concluding remarks.

2. RELATED WORK

Over the last few years, truth finding has become an active research area [21, 3, 5, 24, 9, 11]. Early truth-finding methods either take the *mean* or *median* (for numerical data) or employ the *majority voting* (for categorical data) to predict the truth. These methods treat every source equally and neglect their quality differences [1]. Recent approaches differentiate sources by giving more credit to trustworthy sources and propose solutions for the quality estimation of data sources. TruthFinder [21] alternately computes two measures, the *confidence of fact* (here, facts refer to values) and the *trustworthiness of source*, from each other through an iterative procedure. Pasternack et al. [16] propose *AverageLog*, *Investment* and *PooledInvestment* to avoid overestimating the trustworthiness of those sources that make more claims. Galland et al. [5] propose *Cosine* and *2-Estimates* to incorporate the mutual exclusion between categorical values. In [5], the authors refine the *2-Estimates* algorithm by introducing a new measure, *hardness of fact*, to estimate how hard in obtaining each fact. Truth-finding has also been modeled as optimization problems. The Conflict Resolution on Heterogeneous Data (CRH) framework recently proposed by Li *et al.* [11] models truth-finding as the problem of minimizing the weighted deviation of multi-source inputs from the estimated truth. Yin and Tan [22] employ a different optimization model and propose a semi-supervised solution.

Most above approaches have the disadvantage that a single evaluation result (e.g., the *confidence of fact* of a value) alone cannot indicate whether the value is true, which is also the reason that we have to adapt some of the existing methods in Section 5.2 for MTF. For better interpretation of evaluation results, Bayesian analysis [3] is introduced as a principled approach to the truth-finding problem, which yields explicit probabilistic estimations. Most current Bayesian-based approaches assume a prior distribution of latent variables, such as a uniform distribution over a single type of values (e.g., false values) [3] or distributions of all latent variables [7, 23, 24, 9]. Many of them develop probabilistic graphical models for handling categorical values [24], numerical values [23], ordinal values [9] and knowledge base triples [7]. Waguih *et al.* [20] summarize and experimentally evaluate these truth-finding methods.

Despite these efforts, most existing studies focus on single-truth-finding, yet little attention has been paid to the more general multi-truth-finding problem (MTF). The only work that we are aware of dealing with MTF is the Latent Truth Model (LTM) proposed in [24]. Based on a probabilistic graphical model, LTM makes strong assumptions on the prior distributions of latent variables, rendering the modeled problem intractable and inhibitive to incorporating various considerations. Distinguishing from previous approaches, our approach features an integrated Bayesian model based on a reformulated MTF model. Besides considering the unique features of MTF, our work also differs from the LTM approach [24] in two aspects: i) no assumption on prior distribution of latent variables and ii) new measures for better data source quality estimation. Both the reformulation model and no requirement of prior distribution of latent variables help reduce the computational load, which has been validated in our experimental studies (see Section 5).

[2] Dependent sources are those sources that rely on other sources to provide data, e.g., copiers or aggregators.

3. MULTI-TRUTH-FINDING PROBLEM

3.1 The Problem Model

In general, a multi-truth-finding problem (MTF) involves four basic inputs: i) *data items*, the true values of which are to be discovered, e.g., the author-names of a book, ii) *sources*, which provide values on data items, e.g., a website that publishes the information on books and authors, iii) *values*, e.g., the author-names published by a website, and iv) *mappings* among the above elements, e.g., which websites publish which author's which books.

For each data item, MTF aims at identifying an optimal subset of values from the multi-source inputs to approximate the truth. Multi-truth-finding differs from single-truth-finding in that each source may claim multiple values—instead of a single value—on a single item, and multiple true values may hold on a single item.

Suppose m sources, $\mathcal{S} = \{s_1, s_2, \ldots, s_m\}$, provide values on n items, $\mathcal{O} = \{o_1, o_2, \ldots, o_n\}$. We denote by $\mathcal{S}_i$ the sources that provide values on item o_i, $\mathcal{O}(s_i)$ the items on which a source s_i provides values, and $\mathcal{V}_{ij}$ the values provided by source s_i on item o_j. To describe the mappings between sources and values, we further denote by $\mathcal{S}_i(v)$ the data sources that provide a specific value v on item o_i, and $\mathcal{V}_i(s)$ the values provided by a specific source s on item o_i.

3.2 Reformulating the Problem Model

MTF is inherently difficult and prohibitive to be solved directly. Given a set of possible true values $\mathcal{V}$, any element of the power set of $\mathcal{V}$, instead of any single value of $\mathcal{V}$ in STF, could be the actual truth in MTF. Intuitively, MTF can be first transformed into its single-truth counterparts to be solvable by the existing approaches. However, a direct transformation could excessively expand the problem scale and the unique features of MTF may not be preserved.

To address these problems, we propose to reformulate the MTF model by grouping sources and values based on their mapping relationships over all data items. For ease of illustration, we depict the source-value mappings under different models of MTF with respect to a single data item in Figure 1. Each subfigure shows a bipartite graph/hypergraph that maps sources (or sets of sources) and values (or sets of values) via edges. The three models are as the following:

- The *multi-mapping* model (Figure 1a): A many-to-one mapping between sources and sets of values, which represents the original MTF model as described in Section 3.1.

- The *single-mapping* model (Figure 1b): A many-to-many mapping between sources and values, which represents the result of casting an MTF directly to its single-truth counterparts.

- The *group-mapping* model (Figure 1c): A many-to-many mapping between groups of sources and groups of values, which represents our reformulated model.

Under the single-mapping model (Figure 1b), edges between sources and sets of values can be simply replaced with the edges between sources and individual values (e.g., the three edges between data source s_2 and values v_3, v_4, v_5). Interestingly, the single-truth-finding problem (STF), which is a many-to-one mapping between sources and values on a single item, immediately transforms to the single-mapping model when multiple items are concerned. This explains why the single-mapping model can be directly solved by existing single-truth-finding methods.

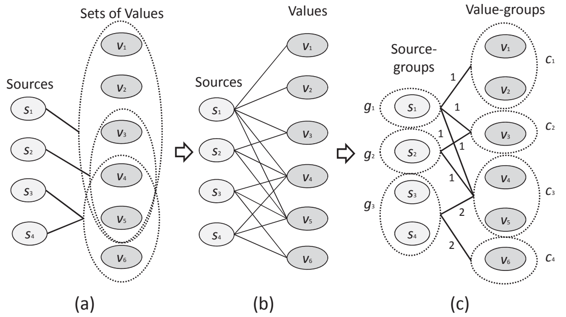


Figure 1: An example illustrating four sources claiming six potential true values under different models of the multi-truth-finding problem.

Though viable, transforming an MTF directly to the single-mapping model tends to result in an exploded problem scale, represented by a multiplied number of nodes in the resulting graph. This could in turn complicate the computation load of the truth-finding methods. As an example, the three nodes in the right side of Figure 1a—which are actually three overlapping sets of values—are decomposed into six nodes in Figure 1b. To reduce the resulting problem scale, instead of decomposing each set into single values, we group the sources (resp., values) that share the same mapping schema in Figure 1b. Each source-group represents the maximum number of sources that claim the same set of values. Similarly, each value-group represents the maximum number of values that are claimed by the same set of sources. As an example, sources s_3 and s_4 in Figure 1c claim the same set of values $\{v_4, v_5, v_6\}$, so s_3 and s_4 are grouped together as a source-group g_3. While v_6 is solely claimed by s_3 and s_4, v_1 and v_2 are claimed by s_1, so, v_1 and v_2 are grouped together as a value group c_1, and v_6 alone as a group c_4. We can see that after the grouping, the node size is reduced from 10 (4:6) in Figure 1b to 7 (3:4) in Figure 1c.

We introduce new concepts of *source-group* and *value-group* to define our reformulated problem model. In particular, we denote by $\mathcal{G}$ the set of source-groups, $\mathcal{C}$ the set of value-groups, $\mathcal{G}_n$ the source-groups that claim values on item o_n, $\mathcal{O}(g_k)$ the items on which a source group g_k claims values, and $\mathcal{C}_{kn}$ the value-groups claimed by the source-group g_k on item o_n. To describe the mapping, we further denote by $\mathcal{G}_n(c)$ the source-groups that claim a specific value-group c on item o_n, and $\mathcal{C}_n(g)$ the value-groups claimed by a specific source-group g on item o_n.

Different from value-groups, each source-group represents the joint strength of all the member sources. To represent this joint strength, we add weights to the edges associated with source-groups in Figure 1c. Given a data item o_n, we define the weight on the edge between source-group g and an associated value-group c as $\omega(g, c) = |g|$, where $c \in \mathcal{C}_n(g)$, $|g|$ is the number of sources contained in g. For the example in Figure 1c, both the edges associated with g_3 should be weighted by 2 because g_3 contains two sources. All the other edges are weighted by 1 because they each contains only one source. After the weighting, each source-group and each value-group will be considered as a single node in the subsequent truth-finding process.

4. THE TRUTH-FINDING MODEL

In this section, we introduce the details of our approach, including the methods on grouping sources and values, the integrated Bayesian framework and the corresponding algorithms. The main notations used in this paper are summarized in Table 1.

4.1 Grouping Sources and Values

Grouping methods aim at reducing the scale of the truth discovery problems. To this end, we expect each group to be as large as possible, to maximally reduce the computation load. Meanwhile, we expect the elements in each group to be as similar as possible, so as to keep the computation simple.

In our approach, we group sources directly based on the multi-mapping model for all data items, which is similar to Figure 1a, but involves multiple data items. We first map all distinct values to a Hash table, and then calculate the sum of hash values regarding each source. If two sums turn out equal, the corresponding sources are further compared with respective to their claimed values. In this way, we gradually assemble similar sources until all sources associated with the same set of values are grouped together. In case the hash values are non-additive, we designate a unique sequence over all the different values, and group those sources that map to the same subsequences. Values are grouped with respect to each data item in three steps based on the resulting source-groups. First, we transform the multi-mapping into the single-mapping model, and then transform the source-value mapping in the single-mapping model into the mapping between source-groups and values. Finally, the values are grouped in the similar way as we group sources. The time complexity of the grouping methods is $O(|\mathcal{S}||\mathcal{V}|)$.

4.2 Integrated Bayesian Model

The Bayesian model estimates the *a posteriori* veracity of values (i.e., latent variables) based on sources' trustworthiness (i.e., model parameters) and sources' reports on potential true values (i.e., observations) by Eq.(1). The sources' trustworthiness can in turn be assessed by the estimated veracity.

$$\mathcal{P}(a(c)|\mathcal{X}) = \frac{\mathcal{P}(\mathcal{X}|a(c))\mathcal{P}(a(c))}{\sum_{a \in \{a(c), \bar{a}(c)\}} \mathcal{P}(\mathcal{X}|a)\mathcal{P}(a)} \tag{1}$$

Both the *a priori* veracity and sources' trustworthiness are manually defined, and the conditional probabilities are calculated by:

$$\mathcal{P}(\mathcal{X}|a) = \prod_{g \in G^{+}(a)} \tau(g) \prod_{g' \in \mathcal{G}^{-}(a)} (1 - \tau(g')) \tag{2}$$

In our approach, we extend the basic Bayesian model by incorporating the following considerations:

- *Degree of claim*, $w(g, c)$, represents the weights on the edges of the bipartite graph (defined in Section 3.2). The concept is the co-product of reducing the problem scale by using the source-grouping method.

- *Confidence score*, $\mu(g, c)$, is used to quantify the impact of mutual exclusive relation between different categorical values on multi-truth-discovery. We develop the corresponding methods to characterize the multi-truth features (to be detailed in Section 4.3).

- *Independence score*, $\mathcal{I}(g, c)$, is specified to quantify the impact of source dependency to multi-truth-discovery. We develop a finer-grained copy detection method to deal with sources with large profiles (to be detailed in Section 4.4).

In addition, since combining positive and negative perspectives can help better distinguish between sources with truth-sensitive and fault-sensitive behavioral features, we use *positive precision* (τ_{pp})—precision on true samples, and *negative precision* (τ_{np})—precision on false samples, to replace τ in Eq.(2). We define the above two measures based on the *veracity score* ($\sigma(c)$) of value-groups as follows:

$$\begin{cases} \tau_{pp}(g) = \frac{\sum_{o_n \in \mathcal{O}} \sum_{c \in \mathcal{C}_n(g)} \sigma(c)}{\sum_{o_n \in \mathcal{O}} |\mathcal{C}_n(g)|} \\ \tau_{np}(g) = \frac{\sum_{o_n \in \mathcal{O}} \sum_{c' \in \mathcal{C}_n(g)} (1 - \sigma(c'))}{\sum_{o_n \in \mathcal{O}} |\mathcal{C}_n \setminus \mathcal{C}_n(g)|} \end{cases} \tag{3}$$

We compute *veracity score* as the truth probability of value-groups using the extended Bayesian model. We find the *degree of claim* naturally resides over quality measures as powers in the Bayesian model and should not be normalized—the Bayesian model calculates the joint effect of sources by multiplying their respective effects, and the multiplication turns into a power function when all sources have equal effect. Indeed, the Bayesian model requires modeling all factors as powers because simple multipliers will be eliminated during calculation. Therefore, we model above parameters as powers over the quality measures in our model. Here, we simply take the product of the different scores to represent their joint effect, but leave more sophisticated combinations of the scores to our future work. For simplicity, we synthesize the parameters into four factors:

$$\begin{cases} f(g, c) = \tau_{pp}(g)^{\omega(g,c)\mathcal{I}(g,c)\mu(g,c)} \\ \mathcal{J}(g, c) = (1 - \tau_{np}(g))^{\omega(g,c)\mathcal{I}(g,c)\mu(g,c)} \\ \hat{f}(g, c) = (1 - \tau_{pp}(g))^{\omega(g,c)\mathcal{I}(g,c)\mu(g,c)} \\ \hat{\mathcal{J}}(g, c) = \tau_{np}(g)^{\omega(g,c)\mathcal{I}(g,c)\mu(g,c)} \end{cases} \tag{4}$$

Given a value-group c, we define the likelihood of $\mathcal{X}$ under different assumptions on the truthfulness[3] of c:

$$\begin{cases} \mathcal{P}(\mathcal{X}|a(c)) = \prod_{g \in \mathcal{G}^{+}(a(c))} f(g, c) \prod_{g' \in \mathcal{G}^{-}(a(c))} \mathcal{J}(g', c) \\ \mathcal{P}(\mathcal{X}|\bar{a}(c)) = \prod_{g' \in \mathcal{G}^{+}(\bar{a}(c))} \hat{\mathcal{J}}(g', c) \prod_{g \in \mathcal{G}^{-}(\bar{a}(c))} \hat{f}(g, c) \end{cases} \tag{5}$$

The source-groups that support or oppose the same assertions should have the following relations:

$$\mathcal{G}^{+}(a(c)) = \mathcal{G}^{-}(\bar{a}(c)), \mathcal{G}^{-}(a(c)) = \mathcal{G}^{+}(\bar{a}(c)) \tag{6}$$

By substituting Eq.(5) into Eq.(1) and adopting Eq.(6), we have:

$$\mathcal{P}(a(c)|\mathcal{X}) = \frac{1}{1 + \frac{\mathcal{P}(\bar{a}(c))}{\mathcal{P}(a(c))} \prod_{g \in \mathcal{G}^{+}(a(c))} \frac{\hat{f}(g,c)}{f(g,c)} \prod_{g' \in \mathcal{G}^{-}(a(c))} \frac{\hat{\mathcal{J}}(g',c)}{\mathcal{J}(g',c)}} \tag{7}$$

4.3 Calculating Confidence Scores

Since most truth-finding methods tend to favor sources with large profiles, incorporating the mutual exclusive relation can significantly neutralize this effect and therefore improve truth discovery accuracy on categorical data. An example of mutual exclusion is that, by claiming *Washington, D.C.* as the capital city of the United States, a source

[3] Truthfulness could be either true or false.

Table 1: Notations used in the paper

Notation	Explanation
s, g	A source (resp., source-group)
v, c	A value (resp., value-group)
$\mathcal{S}$, $\mathcal{G}$	The set of all sources (resp., all source-groups)
$\mathcal{V}$, $\mathcal{C}$	The set of all values (resp., all value-groups)
$\mathcal{G}(c)$, $\mathcal{S}(c)$	The set of all source-groups (resp., sources) that claim a specific value-group
$\mathcal{C}(g)$	The set of all value-groups claimed by a specific source-group
$a(c)$, $\bar{a}(c)$	The assertion that a specific value-group is true (resp., false)
$\mathcal{G}^+(a)$, $\mathcal{G}^-(a)$	The set of all source-groups that supports (resp., oppose) an assertion $a \in \{a(c), \bar{a}(c)\}$
$\omega(g, c)$	The degree of the claims made by a source-group on a value-group
$\mathcal{I}(g, c)$, $\mathcal{I}(s, c)$	The independence score of a source-group (resp., a source) on a value-group
$\mu(g, c)$	The confidence score of a source-group on a value-group
$\tau(g)$	The trustworthiness of a source-group
$\tau_{pp}(g)$, $\tau_{np}(g)$	The positive precision (resp., negative precision) of a source-group
$\sigma(c)$	The veracity (i.e., the probability of being true) of a value-group
ρ	The probability of a source-group copying from other source-groups
ρ_t, ρ_f	The probability of a source-group copying a true (resp. false) value-group from other source-groups
$\mathcal{X}$	The observation of which source-groups claim which value-groups
Ψ_c	The observation that two source-groups claim the same specific value-group

implicates that all other cities are not. Similarly, we can define mutual exclusion between sets of values for MTF. However, traditional truth-finding methods assume that a source always supports or opposes an assertion by its full credit. In fact, in MTF, a claimed value does not strictly reject the unclaimed values because each source could provide only partial true values. We use the *confidence score*, $\mu(g, c)$, to quantify the strength that a source-group supports or opposes an assertion. Similar to the *Kappa coefficient* [6], the idea is to exclude the effect of random guess in determining the strength. More specifically, given a set of value-groups $\mathcal{C}$, if a source-group g claims a subset $\mathcal{C}(g) \subseteq \mathcal{C}$, the *confidence score* of g on each value-group c is calculated as:

$$\mu(g,c) = \begin{cases} \dfrac{1}{|\mathcal{C}(g)|}(1 - \dfrac{1}{|\mathcal{C}|}), & c \in \mathcal{C}(g) \qquad (8a) \\ \dfrac{1}{|\mathcal{C}\backslash\mathcal{C}(g)|}\dfrac{1}{|C|}, & c \in \mathcal{C}\backslash\mathcal{C}(g) \qquad (8b) \end{cases}$$

Based on above definition, by claiming certain value-groups, a source-group supports each claimed value-group and opposes each unclaimed value-group at the same time with the *confidence score*s defined by Eq.(8a) and Eq.(8b), respectively. All the *confidence score*s regarding the same source-group sum up to 1, where each score $\mu(g, c) \in (0, 1]$.

Generally, the *confidence score* has the following interesting properties:

- Given a fixed set of value-groups, the more (resp., less) value-groups a source-group claims, the less (resp., more) confidence the source-group has on the claimed value-groups, and meanwhile the more (resp., less) confidence on the unclaimed value-groups.
- Given a fixed number of value-groups claimed by a source-group, the larger (resp., smaller) the set of value-groups are, the more (resp., less) confidence the source-group has on the claimed (resp., unclaimed) value-groups.

As an example, suppose a source-group claims a subset $\{v_1, v_3\}$ of $\{v_1, v_2, v_3, v_4, v_5\}$, the traditional method obtains the corresponding scores as $\{+1, -1, +1, -1, -1\}$. In contrast, our method would produce $\{\frac{+2}{5}, \frac{-1}{15}, \frac{+2}{5}, \frac{-1}{15}, \frac{-1}{15}\}$[4]. In particular, the values $\frac{2}{5}$ is calculated as $\frac{1}{2} \cdot (1 - \frac{1}{5})$ by using Eq. (8a) and $\frac{1}{15}$ is calculated as $\frac{1}{3} \cdot \frac{1}{5}$ by using Eq. (8b). Compare to the results of the traditional method, our results reflect a differentiation towards source-groups' confidence on the claimed and unclaimed value-groups, i.e., $\frac{2}{5}$ for the claimed and $\frac{1}{15}$ for the unclaimed value-groups, instead of 1 for both types of value-groups. This is important because disclaiming a value is no longer equivalent to disclaiming the value in the MTF's context. Besides, our results implicitly reflect a differentiation towards source-groups of different behavioral features. Following the above example, if a source-group claims another subset $\{v_1, v_3, v_4, v_5\}$, which is closer to the full set, our method would redeem the source-group as being more audacious than being cautious and therefore lower the confidence on the claimed value-groups (meanwhile increase the confidence on the unclaimed value-groups), as manifested by the new results $\{\frac{+1}{5}, \frac{-1}{5}, \frac{+1}{5}, \frac{+1}{5}, \frac{+1}{5}\}$.

4.4 Inferring Source Dependency

Although copying relation has been actively studied recently [3, 14, 19], existing copy detection techniques only calculate a global score for each source [13]. Thus, they can hardly be applied to cases with partial dependence and/or high-order dependence. Especially according to the long tail characteristics [10], a source may have an extremely large profile (e.g., the store *A1Books*, which is a source in the *book-author* dataset, published nearly 700 book records on *www.abebooks.com*). Under such condition, a global score cannot manifest the characteristics of all different parts of the source's data. In contrast, a finer-grained copy detection technique will produce better predictions.

Based on this insight, we introduce a new copy detection method to calculate the *independence score* for each (source-group, value-group) pair. Given such a pair (g, c), we first calculate a score, $\mathcal{I}(s, c)$, for each (source, value-group) pair (s, c), where $s \in g$, and then aggregate the above scores to derive to independence score for (g, c) as follows:

$$\mathcal{I}(g,c) = \frac{\sum_{s \in g} \mathcal{I}(s,c)}{|g|} \qquad (9)$$

where $\mathcal{I}(g, c)$ is the independence score of source-group g on value-group c.

[4] In both results, the positive (resp., negative) sign represents the source-group supports (resp., opposes) the assertion that the corresponding value is true.

4.4.1 Calculating Independence Score

Copying only happens between sources that provide the same value-group. Based on this observation, we propose to calculate $\mathcal{I}(s, c)$ by examining the independence probability of s on every other sources that provide the same value-group c. Given Ψ_c, the observation that two sources provide the same value-group c, we denote by $\perp$ (resp., $\sim$) the independence (resp., copying) relation between sources on c, and $\rightarrow$ the former copies c from the latter. Note that, we omit parameterizing the above notations by c (except Ψ_c) for ease of description. For two arbitrary sources s and s_i ($s \neq s_i$), we have:

$$\begin{cases} \mathcal{P}(s \perp s_i|\Psi_c) + \mathcal{P}(s \sim s_i|\Psi_c) = 1 \\ \mathcal{P}(s_i \rightarrow s|\Psi_c) + \mathcal{P}(s \rightarrow s_i|\Psi_c) = \mathcal{P}(s \sim s_i|\Psi_c)) \end{cases} \tag{10}$$

Given a value-group c, and any source that claims c, $s \in \mathcal{S}(c)$, we define the independence score of s on c as the probability that s never copies c from other sources:

$$\mathcal{I}(s, c) = \prod_{s_i \in \mathcal{S}(c) \wedge s_i \neq s} 1 - \mathcal{P}(s \rightarrow s_i|\Psi_c) \tag{11}$$

We assume equal probability of the two directions of copying, i.e.,

$$\mathcal{P}(s_i \rightarrow s|\Psi_c) = \mathcal{P}(s \rightarrow s_i|\Psi_c) \tag{12}$$

By incorporating Eq.(10) and Eq.(12), we reform Eq.(11) into:

$$\mathcal{I}(s, c) = \prod_{s_i \in \mathcal{S}(c) \wedge s_i \neq s} \frac{1 + \mathcal{P}(s \perp s_i|\Psi_c)}{2} \tag{13}$$

4.4.2 Calculating Independence Probability

To calculate the probability of independence between two sources, we first define the likelihood of Ψ_c under different assumptions on source dependence and the truthfulness of c:

$$\begin{cases} \mathcal{P}(\Psi_c|s_1 \sim s_2, a(c)) = \mathcal{P}(\Psi_c|s_1 \sim s_2, \bar{a}(c)) = 1 \\ \mathcal{P}(\Psi_c|s_1 \perp s_2, a(c)) = \prod_{s \in \{s_1, s_2\}} \tau_{pp}(s) \prod_{s \in \{s_1, s_2\}} \theta \\ \mathcal{P}(\Psi_c|s_1 \perp s_2, \bar{a}(c)) = \prod_{s \in \{s_1, s_2\}} \tau_{np}(s) \prod_{s \in \{s_1, s_2\}} \eta \end{cases} \tag{14}$$

Here, θ=1 - $\tau_{np}(s)$ and η=1 - $\tau_{pp}(s)$. For any assumption $d \in \{s_1 \sim s_2, s_1 \perp s_2\}$, we develop Bayesian formulas to calculate the corresponding probability, where $a \in \{a(c), \bar{a}(c)\}$:

$$\begin{aligned} \mathcal{P}(d|\Psi_c) =& \frac{\mathcal{P}(\Psi_c|d)\mathcal{P}(d)}{\sum_{d'} \mathcal{P}(\Psi_c|d')\mathcal{P}(d')} \\ =& \frac{\sum_a \mathcal{P}(\Psi_c|d, a)\mathcal{P}(d|a)\mathcal{P}(a)}{\sum_{d'} \sum_{a'} \mathcal{P}(\Psi_c|d', a')\mathcal{P}(d'|a')\mathcal{P}(a')} \end{aligned} \tag{15}$$

In our approach, we distinguish between two types of copiers, namely *blind copiers* and *smart copiers*. The blind copiers assume independence between the veracity of values and sources' probability of copying, i.e., $\mathcal{P}(d|a) = \mathcal{P}(d)$. We can thereby rewrite Eq.(15) to:

$$\mathcal{P}(d|\Psi_c) = \frac{\mathcal{P}(d) \sum_a \mathcal{P}(\Psi_c|d, a)\mathcal{P}(a)}{\sum_{d'} \mathcal{P}(d') \sum_{a'} \mathcal{P}(\Psi_c|d', a')\mathcal{P}(a')} \tag{16}$$

Since blind copiers have no bias on copying true/false values, we define a single *copying probability* ρ for all sources and on all value-groups:

$$\begin{cases} (\mathcal{P}(s_1 \sim s_2) = \mathcal{P}(s_1 \rightarrow s_2) + \mathcal{P}(s_2 \rightarrow s_1) = 2\rho \\ \mathcal{P}(s_1 \perp s_2) = 1 - 2\rho \end{cases} \tag{17}$$

By substituting Eq.(14) and Eq.(17) into Eq.(16), we get:

$$\mathcal{P}(s_1 \perp s_2|\Psi_c) = \frac{(1 - 2\rho)\mathcal{P}_{sum}}{2\rho + (1 - 2\rho)\mathcal{P}_{sum}} \tag{18}$$

where $\mathcal{P}_{sum}$ denotes the sum term in the numerator of Eq.(16):

$$\begin{aligned} \mathcal{P}_{sum} =& \mathcal{P}(a(c))(\tau_{pp}(s_1)\tau_{pp}(s_2) + (1 - \tau_{np}(s_1))(1 - \tau_{np}(s_2))) \\ &+ \mathcal{P}(\bar{a}(c))(\tau_{np}(s_1)\tau_{np}(s_2) + (1 - \tau_{pp}(s_1))(1 - \tau_{pp}(s_2))) \end{aligned} \tag{19}$$

Without prior knowledge, we can initialize veracity as:

$$\forall c \in \mathcal{C},\ \mathcal{P}(a(c)) = \mathcal{P}(\bar{a}(c)) = 0.5$$

On the other hand, the smart copiers have some "smartness" that they are more likely to copy true value-groups than false value-groups. We define different conditional probabilities for the two cases to reflect the "smartness":

$$\mathcal{P}(s_1 \sim s_2|a(c)) = 2\rho_t,\ \mathcal{P}(s_1 \sim s_2|\bar{a}(c)) = 2\rho_f \tag{20}$$

It can be inferred from the above equations that:

$$\mathcal{P}(s_1 \perp s_2|a(c)) = 1 - 2\rho_t,\ \mathcal{P}(s_1 \perp s_2|\bar{a}(c)) = 1 - 2\rho_f \tag{21}$$

For copiers to be "smart", the probability of a source copying a true value-group should be larger than the probability of copying a false value-group, i.e., $\rho_t > \rho_f$. By substituting Eq.(14)(20)(21) into Eq.(15), we get:

$$\mathcal{P}(s_1 \perp s_2|\Psi_c) = \frac{\mathcal{P}_{over}}{2\rho_t\mathcal{P}(a(c)) + 2\rho_f\mathcal{P}(\bar{a}(c)) + \mathcal{P}_{over}} \tag{22}$$

where $\mathcal{P}_{over}$ denotes the numerator in Eq.(15):

$$\begin{aligned} \mathcal{P}_{over} =& (1 - 2\rho_t)\mathcal{P}(a(c)) \\ & (\tau_{pp}(s_1)\tau_{pp}(s_2) + (1 - \tau_{np}(s_1))(1 - \tau_{np}(s_2))) \\ &+ (1 - 2\rho_f)(1 - \mathcal{P}(a(c))) \\ & (\tau_{np}(s_1)\tau_{np}(s_2) + (1 - \tau_{pp}(s_1))(1 - \tau_{pp}(s_2))) \end{aligned}$$

Because it is critical for smart copiers to acquire some prior knowledge in order to be "smart", we update the prior probability with the latest estimation of veracity scores after each cycle of the iteration:

$$\forall c \in \mathcal{C},\ \mathcal{P}(a(c)) \leftarrow \sigma(c), \mathcal{P}(\bar{a}(c)) \leftarrow 1 - \sigma(c) \tag{23}$$

This ensures that the smart copiers' perception on values' veracity keeps evolving with the truth-finding process.

4.5 The Algorithm

Various algorithms, such as the iteration algorithm [21, 3, 5, 17] and the Expectation Maximization (EM) algorithm [23, 2], can be applied to solve our model. Both algorithms belong to the category of *coordinate ascent algorithms*, which differ in the methods used for estimating the quality of data sources. In particular, the former defines linear or nonlinear functions to calculate sources' quality, while the latter infers sources' quality by maximizing the (lower bound of the logarithmic) likelihood of observations over all source-claimed values.

Here we present an iteration algorithm for our integrated Bayesian model, but omit the description of the EM algorithm, which is only slightly different from [18], due to the limited space. For the ease of illustration, we use a single notation to represent the copying probabilities of the two

```
Algorithm 1: Iterative Multi-Truth-Finding
Input: data items $\mathcal{O}$, value-groups $\mathcal{C} = \{\mathcal{C}_n | o_n \in \mathcal{O}\}$,
       source-groups $\mathcal{G}$, the mapping between source-groups
       and value-groups $\mathcal{G}_n(c)$ and $\mathcal{C}_n(g)$
Output: $\{v | v \in c \wedge \sigma(c) \geq 0.5\}$
1  $\rho \leftarrow$ default values;
2  foreach $g \in \mathcal{G}$ do
3      $\{\tau_{pp}(g), \tau_{np}(g)\} \leftarrow$ default values;
4  foreach $o_n \in \mathcal{O}$ and $c \in \mathcal{C}_n$ do
5      $\sigma(c) \leftarrow$ default values;
6  foreach $o_n \in \mathcal{O}$, $g \in \mathcal{G}_n(c)$, $c \in \mathcal{C}_n(g)$ do
7      compute $w_n(g, c)$, $u_n(g, c)$;
8  repeat
9      foreach $o_n \in \mathcal{O}$ and $c \in \mathcal{C}_n$ do
10         foreach $g \in \mathcal{G}_n(c)$ do
11             $f_n(g, c)$,$\hat{f}(g, c)$,$J_n(g, c)$, $\hat{J}_n(g, c) \leftarrow$Eq.(4);
12             $I_n(g, c) \leftarrow$Eq.(9) ;
13         $\sigma(c) \leftarrow$Eq.(7);
14     foreach $g \in \mathcal{G}$ do
15         $\tau_{pp}(g)$, $\tau_{np}(g) \leftarrow$Eq.(3);
16 until convergence;
```

types of copiers:

$$\rho = \begin{cases} \{\rho\}, & \textit{for blind copiers} \\ \{\rho_t, \rho_f\}, & \textit{for smart copiers} \end{cases}$$

The detailed procedure is described in Algorithm 1. In the initialization phase (Lines 1-7), the copying probabilities are defined *a priori* (Line 1). Sources' quality and values' veracity are initialized with default values (Lines 2-5). The algorithm then computes the *degree of claim* and *confidence score* for each pair of source-group and value-group (Lines 6-7). Both parameters and the copying probabilities remain unchanged until the algorithm terminates. For each cycle of the iteration (Line 8-16), the algorithm calculates the veracity scores (Lines 9-13) and sources' quality (Lines 14-15) in turn. For each data item, the veracity scores are calculated in two steps: i) calculates the synthesized factors (as defined by Eq.(4)) and *independence score* for each pair of source-group and value-group (Lines 10-12) and ii) updates the veracity scores for each value-group (Line 13). The iteration terminates when the algorithm converges (i.e., the algorithm's judgment on the truthfulness of all values remains unchanged for certain consecutive cycles) (Line 16). Logarithms are used in calculating the multiplication of small decimals to ensure accuracy.

5. EXPERIMENTS

In this section, we report our experimental studies on the comparison of our approach with the state-of-the-art algorithms and the impact on the performance of our approach of different key aspects in the Bayesian model.

5.1 The Datasets

We used three real-world datasets in our experiments. The *book-author dataset* [21] contains 33,971 book-author records crawled from *www.abebooks.com*. The records of the website are contributed by numerous book stores (i.e., sources), where each record represents a book store's claim on the author(s) of a book. We removed the invalid and duplicated records. To make the problem more challenging, we also excluded the records with only minor conflicts (i.e., the records related to those books on which less than two distinct lists of author-names are provided). Finally, we obtained 12,623 distinctive claims describing 649 sources (i.e., websites) that provide author-names on 664 books. On average, each book has 3.2 authors. The ground truth provided for the original dataset is used as gold standard.

Table 2: Evaluation of the major sources in the *movie-director* dataset.

Source	Record number	Positive precision	Negative precision
hkmdb.com	5265	0.91	0.87
rottentomatoes.com	4950	0.92	0.95
mrqe.com	4931	0.98	0.93
nowrunning.com	3433	0.83	0.81
imdb.com	2592	0.98	0.94
moviefone.com	2529	0.92	0.86
dvdmoviemenus.com	2208	0.87	0.69
abc.net.au	1563	0.99	0.84
nollywoodreinvented.com	1071	0.95	0.80
hoyts.com.au	1066	0.96	0.93

The *parent-children dataset* [16] contains 11,099,730 records about people's birth and death dates, the names of their parents/children and spouses, edited by different users (i.e., sources) on Wikipedia. We particularly extracted the records on the parent-children relations from this dataset. After eliminating the duplicates, we finally obtained 55,259 users claiming children for 2,579 persons. In the resulting dataset, each person has on average 2.45 children. We used the latest editing records as the ground truth.

We prepared the third dataset, the *movie-director dataset*, by crawling 33,194 records from 16 movie websites. We removed redundant records and finally obtained 6,402 movies, each on average having 1.2 directors. We sampled 200 movies and extracted their director information from *citwf.com* as the ground truth. Table 2 shows the top ten websites that provide the most records, with their quality values obtained by one of our methods MBM (see Section 5.2 for details). It should be noted that most datasets used in previous works for categorical truth discovery [12, 20] are not suitable for our multi-truth-finding problem. The three real datasets used in our work are comparable to those datasets in size.

5.2 Baselines and Metrics

We compared our approach with the following methods, which were modified, if necessary, to incorporate mutual exclusion.

- *Majority Voting.* This method regards a value as true if the proportion of the sources that claim the value exceeds a certain threshold.
- *Sums (Hubs and Authorities)* [8], *Average-Log* [16]. Both methods compute the total trustworthiness of all sources that claim and disclaim a value separately, and recognize the value as true if the former is larger than the latter.
- *TruthFinder* [21], *2-Estimates* [5], and *LTM* [24]. The three methods can be directly applied without modification, which recognize a value as true if its veracity score exceeds 0.5.

It should be noted that we excluded the comparison with several methods that are inapplicable to the multi-truth-finding problem. For example, the algorithms in [3] cannot be applied to our problem because they all assume the number of false values as a prior knowledge. The approach

Table 3: Comparison of different algorithms on the three datasets: the best and second best performance values are in bold; both *precision* and *recall* are in the range of [0,1].

Method	Book-author dataset			Parent-children dataset			Movie-director dataset		
	Precision	Recall	Time(s)	Precision	Recall	Time(s)	Precision	Recall	Time(s)
Majority voting	0.88	0.62	**0.03**	**1.00**	0.52	**1.25**	**1.00**	0.74	**0.07**
Sums	0.69	0.49	**0.10**	0.86	0.59	3.20	0.88	0.64	0.22
Average-Log	0.70	0.38	0.13	0.89	0.79	4.08	0.89	0.87	0.24
TruthFinder	0.73	0.80	**0.10**	0.86	**0.89**	**2.89**	0.87	**0.88**	0.27
2-Estimates	0.79	0.65	0.12	0.92	0.62	3.44	0.89	0.77	**0.21**
LTM	0.84	0.78	0.86	0.93	0.82	39.2	0.83	0.83	1.84
MBM	0.78	**0.87**	0.16	**0.97**	**0.87**	3.19	0.89	0.87	0.22
MBM-C	**0.88**	0.85	0.50	0.94	0.85	19.1	**0.93**	**0.92**	1.28
MBM-EM	**0.86**	**0.91**	0.26	0.92	0.84	6.87	0.92	**0.88**	0.72

in [16] requires normalizing the veracity of values, which is infeasible for the multi-truth-finding problem. Finally, the methods in [23, 11] focus on handling heterogeneous data, while our approach is proposed specially for categorical data.

To ensure fair comparisons, we first ran a series of experiments to decide the optimal parameter settings for the baseline methods. Since the parameter tuning for our methods are relatively more complicated, we simply used a generic parameter settings for all datasets, i.e., the copying probabilities of blind copiers ρ=0.8 and for smart copiers, ρ_t=0.85 and ρ_f=0.7. The initial source quality values do not usually affect the experimental results as long as they are not unreasonably large or small (as indicated in our experiments in Section 5.3.2), so we just initialized them as $\tau_{pp}(g)$=0.8 and $\tau_{np}(g)$=0.7.

To evaluate our approach under different implementations, we derived three variants of our approach:

- *MBM*: our (**M**ulti-truth) **B**ayesian **M**odel that adopts the grouping method and the new mutual exclusion definition.
- *MBM-C*: a variant of MBM that additionally incorporates our **C**opy detection method for blind copiers[5].
- *MBM-EM*: a variant of MBM that estimates sources' quality by performing the Maximum Likelihood Estimation using the **EM** algorithm, instead of Eq.(3).

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

5.3 The Results

5.3.1 Comparison of Truth-Finding Methods

Table 3 shows the performance of different algorithms on the three datasets in terms of *precision*, *recall*, and *computation time*. The computation time of our algorithms includes the time spent for both problem reformulation and Bayesian truth discovery. However, the results show the time spent on reformulation is minor when compared to that of main truth discovery process. Our three algorithms consistently achieved the best precision and recall among all the compared methods, except the *majority voting* which always achieved the best precision (in those cases, our algorithms still yielded the second best results). All the algorithms achieved lower precision on the *book-author* dataset due to the elimination of the records with minor conflicts.

The *majority voting* achieved comparatively low recall (nearly always the lowest) on all datasets. This is because most sources tend to provide only a minor proportion of the entire truth. So when tuning the sources' trustworthiness as the prior parameters, only the precision of the method is optimized. Despite the low recall, the majority voting achieved nearly perfect precision—except on the *book-author* dataset where the approach is inapplicable. This may imply that the majority voting method is better used for generating the ground truth for semi-supervised truth-finding approaches, rather than for solving MTF, unless more comprehensive quality measures are considered in evaluating the sources.

Besides the majority voting, both LTM and 2-Estimates showed higher precision than the other baselines. All baselines except TruthFinder considered the mutual exclusive relation. However, these methods achieved lower recall when compared to TruthFinder or our methods. They identified only a small proportion of true values. This may be due to their neglect of the possibility of random guess in considering sources' claims—as opposed to the definition of mutual exclusive relation in our approach. This should explain why our methods achieve better recall than those methods. It should be noted that TruthFinder achieved better recall yet generally lower precision than the other baselines, which may attribute to its overestimation of veracity scores.

As for the efficiency, our MBM and all the baselines—except LTM—had comparable computation time on the three datasets. LTM and MBM-C always demanded the longest computation time. While the efficiency of LTM depends on the problem scale, MBM-C is more sensitive to datasets. Specially, MBM-C achieved significantly better efficiency than LTM on the *parent-children* dataset, because of the many source-groups and value-groups in this dataset. Overall, our three methods showed no significant difference in their truth-finding quality. However, MBM-C exhibited less stable performance, depending on the underlying dependence among sources in the datasets. MBM-EM always ranked in the middle of the three in term of efficiency.

5.3.2 Impact of Different Concerns

We also studied the impact of different aspects to our methods and report the findings in this section.

Grouping of sources/values. We exploited our methods to discover various source-groups and value-groups in the three datasets. Table 4 shows examples of three source-groups found in the *book-author* dataset. In the first example, six sources claim the same two authors for a book. In the third example, two sources claim the same author for each of the ten books. By grouping the sources , the number of sources in the three examples was reduced from 10

[5] We only used blind copiers for the comparison because smart copiers tend to produce similar results. They will be specially compared via experiments in Section 5.3.2.

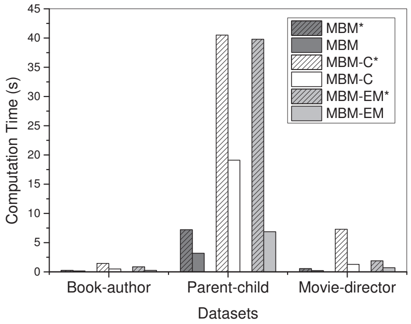


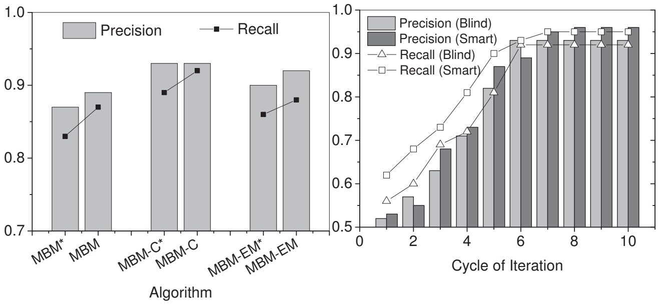
Figure 2: Performance comparison of the proposed algorithms between using and not using the grouping methods. The algorithms marked by asterisk are those without using the grouping methods.

to 3. After the grouping, the total number of sources (or joint sources) in the *book-author dataset* is reduced from 4,264 to 3,874. We found the *author-book dataset* contained more source-groups, while the *movie-director dataset* contains more value-groups. The *parent-children dataset* contains large numbers of both types of groups. A comparison of the three algorithms between using and not using the grouping methods (Figure 2) demonstrates the effectiveness of the grouping methods.

Mutual exclusion. In our datasets, each item has on average 1 to 4 different values, so the confidence scores stay in the range of $(0.08, 0.75)$ (calculated by Eq.(8a) and Eq.(8b)). To examine the effect of our defined confidence scores, we implemented our methods based on the traditional definition and our new definition of mutual exclusive relation, respectively, and compared the results. Figure 3a shows the comparison on the *movie-director* dataset, which demonstrates that our definition almost always brings better precision and recall. The results on the other two datasets are similar.

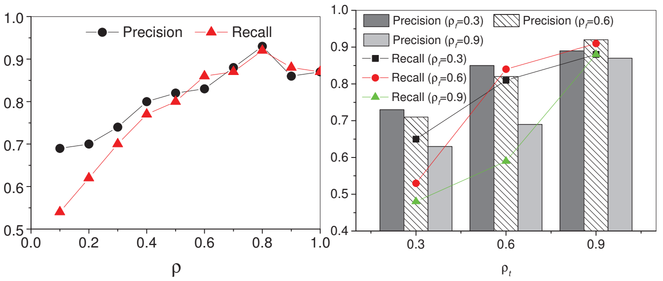
Table 4: An example of three source-groups in the *book-author* dataset.

Source (Website)	Item (ISBN)	Value (Authors)
bookscorner1 Jerome McCarthy Mybooklocator “Rare Finds Books, Music, Etc.” Twice Read Books Shadow Books	0201489163	Campbell David Campbell Mary
bookmac Allen Williams Books	9780335216369	Alyson Simpson Angela Thomas Asha Jennifer Len Unsworth
Usedbooks123 Free Shipping Books	0201308207 0201325705 0201379538 0201379619 0201489295 0201489805 0201633957 0201657643 0201709147 0201711141	Denning Dorothy E Brent Callaghan Black Daryl P Jeffrey Rule S F John Lescher J Smedinghoff Thomas Bradner Scott Sharma Vivek Box Don Aviel D Rubin



(a) (b)

Figure 3: (a) Performance comparison of the proposed algorithms between using the traditional definition and using our definition of mutual exclusion on the *movie-direct* dataset: the algorithms marked by asterisk are those adopting the traditional definition. (b) Performance comparison of MBM-C between using the blind copiers and using the smart copiers. Both blind copiers and smart copiers were configured with their optimal parameter settings and ran a fixed number of iterations, i.e., 10, regardless if they converge.



(a) (b)

Figure 4: (a) Performance of MBM-C under varying copying probability of blind copiers, i.e., ρ. (b) Performance of MBM-C under varying copying probabilities of smart copiers, i.e., ρ_t and ρ_f.

Blind and smart copiers. We investigated the effect of incorporating copy detection by comparing MBM and MBM-C in the experiments. The results showed improved precision and recall of our approach by incorporating copy detection methods. We further studied the performance of our methods using blind copiers and smart copiers, respectively. We observed that using smart copiers led to slightly slower convergence but better results on the *movie-director* dataset (Figure 3b). As the copying probability grew, we observed an increase in both precision and recall of the methods using blind copiers on the movie-director dataset, until the probability became close to 0.8 (Figure 4a). Smart copiers also showed similar features (Figure 4b) on the movie-director dataset (Figure 4b). It is worth noting that, *recall* had some robustness on ρ_f. At certain points, increasing ρ_f could even yield a higher *recall*. The impact of initial parameters were similar for the other datasets.

Comprehensive source quality. We varied the initial values of source quality measures for our methods and observed similar results on all three datasets. This indicates that our approach is insensitive to the initial assumptions of source quality (as long as the initial values are not infeasible large or small, such as equal to one, or close to zero). Compared to the traditional measures, our source quality measures incurred similar computation time but higher recall on larger datasets (e.g., the *parent-children* dataset and the *movie-director* dataset), while the advantages on smaller datasets (e.g., the *book-author* dataset) was not obvious.

5.4 Discussion

In this section, we briefly review the important concepts incorporated in our approach via the *hardness of fact* to better understand the experimental results. The *hardness of fact* was first proposed in [5] to quantify the difficulty in determining the truthfulness of a value. It is used by paying the most trust on the sources that claim a more difficult value (which has a higher *hardness of fact*). We find that both the *smart copiers* and our mutual exclusion definition can be interpreted or inferred from the concept of *hardness of fact* in evaluating the sources. In particular, a smart copier prefers copying the values with higher veracity. Those values are usually claimed by more sources. In defining a higher probability of copying, the smart copier actually dampens the effect of those sources which jointly claim values with many other sources. This is exactly the effect of considering the hardness of fact in the truth-finding process. As for our proposed mutual exclusion definition, a claimed value would receive a higher confidence score if given a larger number of distinct values on a specific item. This can also be interpreted from the *hardness of fact*. Since it is more difficult to identify a true value from a larger set of different values, once a value is identified as true, the value should be more trusted based on the philosophy of the *hardness of fact*.

6. CONCLUSION

In this paper, we address the problem of discovering multiple true values from the multi-source data, which has rarely been studied in the previous works. We propose an integrated Bayesian approach, which comprehensively incorporates novel methods on three key aspects that characterize the multi-truth-finding problem (MTF), namely source-value mapping, mutual exclusive relation, and source dependency, to better solve the problem. In particular, we leverage the unique mapping features of MTF to reformulate the problem model in order to reduce the problem scale. We develop a new definition of mutual exclusion to reflect the inter-value implication under the MTF's context and a finer-grained copy detection method to cope with sources with large profiles. Experimental studies on three real-world datasets demonstrate the effectiveness of our approach. Our future work will focus on investigating more comprehensive ways for solving the MTF, e.g., by identifying and integrating more aspects to enhance the Bayesian model.

7. REFERENCES

[1] J. Bleiholder and F. Naumann. Conflict handling strategies in an integrated information system. In *IIWeb, held together with WWW*, 2006.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977.

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

[4] X. L. Dong and D. Srivastava. Big data integration. In *ICDE*, pages 1245–1248, 2013.

[5] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *WSDM*, pages 131–140, 2010.

[6] K. L. Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[7] G. Kasneci, J. V. Gael, D. Stern, and T. Graepel. Cobayes: bayesian knowledge corroboration with assessors of unknown areas of expertise. In *WSDM*, pages 465–474, 2011.

[8] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *JACM*, 46(5):604–632, 1999.

[9] B. Lakshminarayanan and Y. W. Teh. Inferring ground truth from multi-annotator ordinal data: a probabilistic approach. *arXiv preprint arXiv:1305.0015*, 2013.

[10] 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. *PVLDB*, 8(4), 2014.

[11] 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 *SIGMOD*, pages 1187–1198, 2014.

[12] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava. Truth finding on the deep web: is the problem solved? *PVLDB*, 6(2):97–108, 2012.

[13] X. Li, X. L. Dong, K. B. Lyons, W. Meng, and D. Srivastava. Scaling up copy detection. In *ICDE*, 2015.

[14] X. Liu, X. L. Dong, B. C. Ooi, and D. Srivastava. Online data fusion. *PVLDB*, 4(11):932–943, 2011.

[15] A. Pal, V. Rastogi, A. Machanavajjhala, and P. Bohannon. Information integration over time in unreliable and uncertain environments. In *WWW*, pages 789–798, 2012.

[16] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *COLING*, pages 877–885, 2010.

[17] J. Pasternack and D. Roth. Making better informed trust decisions with generalized fact-finding. In *IJCAI*, pages 2324–2329, 2011.

[18] J. Pasternack and D. Roth. Latent credibility analysis. In *WWW*, pages 1009–1020, 2013.

[19] R. Pochampally, A. D. Sarma, X. L. Dong, A. Meliou, and D. Srivastava. Fusing data with correlations. In *SIGMOD*, pages 433–444, 2014.

[20] D. A. Waguih and L. Berti-Equille. Truth discovery algorithms: An experimental evaluation. *arXiv preprint arXiv:1409.6428*, 2014.

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

[22] X. Yin and W. Tan. Semi-supervised truth discovery. In *WWW*, pages 217–226, 2011.

[23] B. Zhao and J. Han. A probabilistic model for estimating real-valued truth from conflicting sources. In *QDB*, 2012.

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