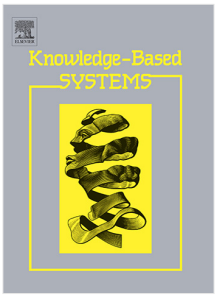
Accepted Manuscript



Truth finding by reliability estimation on inconsistent entities for heterogeneous data sets

Hui Tian, Wenwen Sheng, Hong Shen, Can Wang

PII: S0950-7051(19)30303-X
DOI: https://doi.org/10.1016/j.knosys.2019.06.036
Reference: KNOSYS 4828

To appear in: *Knowledge-Based Systems*

Received date : 30 March 2017
Revised date : 27 June 2019
Accepted date : 29 June 2019

Please cite this article as: H. Tian, W. Sheng, H. Shen et al., Truth finding by reliability estimation on inconsistent entities for heterogeneous data sets, *Knowledge-Based Systems* (2019), https://doi.org/10.1016/j.knosys.2019.06.036

Truth Finding by Reliability Estimation on Inconsistent Entities for Heterogeneous Data Sets

Hui Tian[a], Wenwen Sheng[b], Hong Shen[b,c] and Can Wang[a]
[a]School of Information and Communication Technology, Griffith University, Australia
[b]School of Information Science and Technology, Sun Yat-Sen University, China
[c]School of Computer Science, University of Adelaide, Australia
Email: hui.tian@griffith.edu.au, 771920866@qq.com, hongsh01@gmail.com, can.wang@griffith.edu.au

***Abstract*—An important task in big data integration is to derive accurate data records from noisy and conflicting values collected from multiple sources. Most existing truth finding methods assume that the reliability is consistent on the whole data set, ignoring the fact that different attributes, objects and object groups may have different reliabilities even wrt the same source. These reliability differences are caused by the hardness differences in obtaining attribute values, non-uniform updates to objects and the differences in group privileges. This paper addresses the problem how to compute truths by effectively estimating the reliabilities of attributes, objects and object groups in a multi-source heterogeneous data environment. We first propose an optimization framework TFAR, its implementation and Lagrangian duality solution for Truth Finding by Attribute Reliability estimation. We then present a Bayesian probabilistic graphical model TFOR and an inference algorithm applying Collapsed Gibbs Sampling for Truth Finding by Object Reliability estimation. Finally we give an optimization framework TFGR and its implementation for Truth Finding by Group Reliability estimation. All these models lead to a more accurate estimation of the respective attribute, object and object group reliabilities, which in turn can achieve a better accuracy in inferring the truths. Experimental results on both real data and synthetic data show that our methods have better performance than the state-of-art truth discovery methods.**

***Index Terms*—Truth finding, attribute reliability, object reliability, group reliability, entity hardness, probability graphical model.**

I. Introduction

With the rapid developments of big data and smart city, the need to integrate the true values on heterogeneous data observed from multiple sources together is becoming an urgent task because of the increasing unreliability in object data and observation sources. Reliability inconsistency exists widely in different levels and dimensions. First, apparently observations from different sources for an object may differ from each other due to the difference in data capture ability of the sources, resulting in a many-to-many relationship among Source-Value-Object as illustrated in Figure 1. Moreover, reliabilities of different attributes of an object set wrt the same source may also be different because of the observation hardness differences of the attributes wrt the source (e.g. an RFID reader may have 0.99 reliability for bar-code but only 0.1 for TID). Similarly in an orthogonal dimension, different objects (records) may also carry different reliabilities wrt the same source because of their differences in data entry and maintenance (e.g. the records

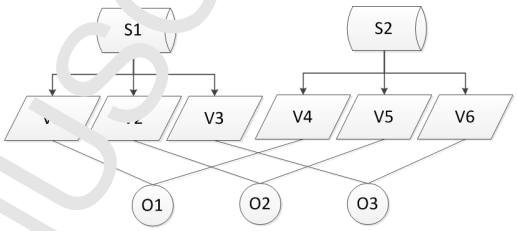


Figure 1. The many-to-many relationship among sources, objects and values

updated frequently may have a higher reliability than those updated infrequently). Finally, we also observe that object reliability is consistent within a group if we divide objects into groups such that all objects in the same group have the same reliability. Examples of group reliability includes privilege groups for online services and user groups in social networks. These reliability inconsistencies will result in source data conflicts and increase the hardness for obtaining the truths for objects.

For truth finding from conflicting data, most existing methods [13], [14], [25] based on majority voting and mean computation for categorical and continuous data respectively took no consideration of source reliabilities and unrealistically treated all observations from all sources equally. Voting selects the majority claims among all the observations as the truth, while mean computation takes the mean of all observations as the truth.

When taking into account of source reliabilities, different truth discovery methods have been proposed [7], [32], [21], [8], [19], [20], all aimed to utilize some sort of specifications about the sources and applied the same basic heuristic idea: a claim is likely to be true if it is provided by trustworthy sources (especially if by many of them) and a source is trustworthy if most its claims are true. Based on this idea, most methods attempted to assign larger weight to reliable sources as they are more important when inferring the truths. These methods however applied the same source reliability to all attributes for each source and are hence unable to distinguish the quality of observations to different attributes from the same source.

We use an example in Table 1 Quiz answers to explain these concepts. *In the data sources shown in Table 1, if we only deal with concrete and continuous data types, the Material*

Object	Digital Analysis	Logical A	Material A
Question 1	8	B	picture11
Question 2	12	B	picture12
Question 3	14	A	picture13

(a) Susan database

Object	Digital Analysis	Logical A	Material A
Question 1	9	A	picture21
Question 2	12	B	picture22
Question 3	13	C	picture23

(b) Mike database

Object	Digital Analysis	Logical A	Material A
Question 1	8	A	picture31
Question 2	12	C	picture32
Question 3	11	C	picture33

(c) Leo database

Table I: Quiz answers of Susan, Mike and Leo

Object	Digital Analysis	Logical A	Material A
Question 1	8	C	picture1
Question 2	14	B	picture2
Question 3	11	A	picture3

Table II: Ground Truth of Quiz

attribute cannot be processed. If we use the source reliability, the reliability degrees of Source 1 (Susan database) and Source 3 (Leo database) are approximate. Nevertheless, Source 1 is more accurate in Logical Analysis attribute and Source 3 is more accurate in Digital Analysis attribute. The answers to Question 3 in Digital Analysis are different from each other, which increases the hardness to get the truth. So the attributes that get answers for harder questions should have a higher reliability and for easier questions a lower reliability.

Existing methods ignored the fact that the same source's reliability may vary significantly among different attributes or objects (records). This motivates our work of this paper to investigate more effective methods for truth finding by reliability estimation on heterogeneous data. We first propose an optimization model TFAR, Truth Finding by Attribute Reliability estimation, to infer the truths by estimating the reliabilities of heterogeneous attributes and the hardness of attribute observation. We obtain a solution for computing an optimal attribute weight (reliability) assignment that minimizes the total deviation between the truths and the observed values. Then we propose a Truth Finding by Object Reliability estimation model (TFOR) using a Bayesian probabilistic graphical model to infer the object reliabilities and truths. We formulate the derivation of the model's parameters as a Maximum Likelihood Estimation problem and apply Collapsed Gibbs Sampling to jointly infer the object reliabilities and truths. Finally we propose another optimization model TFGR for Truth Finding by Group Reliability Estimation to detect trustworthy claims from conflicting observations by estimating the (object) group reliability for the given group properties. We obtain its solution by minimizing the overall weighted deviation between inferred truths in the i-th time (iteration of the deductive procedure) and the source observations to find the final truths. The above three models achieve a more accurate fine-grained source reliability estimation on attributes, objects and object groups respectively.

In our experimental evaluation, we show that our methods outperform the state-of-the-art truth-finding baselines that considered neither attribute reliability differences among all attributes nor object reliability differences among different objects for a source.

The main contributions of this paper are the proposed three mathematical models with their detailed implementation algorithms and solutions to solve the reliability conflict resolution problem for truth finding at attribute, object and object group three levels respectively, as summarized below:

- We propose a general optimization framework for truth finding on inconsistent attribute reliabilities by taking attribute weights and fact hardness into consideration.
- We propose a probabilistic graphical model for truth finding on inconsistent object reliabilities by incorporating quality measurement into object reliability.
- We propose a general optimization framework for truth finding on inconsistent object group reliabilities by iteratively updating group weights.
- We empirically show that our models outperform the existing methods for conflict resolution with three real-world datasets, which demonstrates the importance of taking into consideration reliability differences among attributes, objects and object groups for truth finding on heterogeneous data.

The reminder of this paper is organized as follows: In section 2 we review the related work. Our proposed models and algorithms are introduced in Section 3, Section 4 and Section 5. Section 6 presents the evaluation results. Section 7 concludes the paper.

II. Related Work

The truth finding (conflict resolution) problem was first studied by Yin et al. [31] who proposed a TRUTHFINDER method to iteratively infer the truth values and source quality, and it has now been extensively studied. Existing work can be classified according to the specifications used to measure the source reliability.

Data source specification. The source selection problem identifies the subset of sources that maximizes the profit from integration. Rekatsinas et at defined a set of time-dependent metrics to characterize the quality of integrated data [22]. Dong et al. proposed an approaches of applying Bayesian analysis to decide dependence between sources [3] and select a subset of sources before integration to balance the quality of integrated data and integrated cost [4]. Li et al. studied the long-tail phenomenon in a task (i.e. only a few sources make many claims) and proposed a confidence-aware truth discovery method to estimate the source reliability by considering the confidence interval of the estimation [11].

Observation specification. Wang et al. proposed an approximate truth discovery approach which divides sources and values into groups according to a user specified approximation

criterion, and uses the groups for efficient inter-value influence computation to improve the accuracy [28]. Shi et al. proposed a probabilistic graphical model incorporating silent rate, false spoken rate and true spoken rate three measures to simultaneously infer the truth as well as source quality without any priori training involving ground truth answers [36]. Qi et al. proposed an optimization framework to resolve conflicts among multiple sources of heterogeneous data, where truths and source reliability are defined as two sets of unknown variables [12]. Zhao et al. proposed a probabilistic graphical model that can automatically infer the true records and source quality without supervision, which leverages a generative process for modelling two types of errors from two different aspects of source quality [34]. Zhao and Han proposed a truth-finding method specially designed for handling numerical data based on Bayesian probabilistic modeling on the dependencies among source quality, truth, and claimed values [33].

Crowdsourcing specification. Ma et al. proposed a fine grained truth discovery model for the task of aggregating conflicting data collected from multiple users/sources [18]. Wang et al. addressed the challenge of truth discovery from noisy social sensing data on binary measurements and gave the first optimal solution [27]. Wang et at presented a streaming approach to solve the truth estimation problem in crowdsourcing applications [26]. Whitehill et al. presented a probabilistic model to simultaneously infer the label, expertise and difficulty of an image [30].

With the development of big data analytics, study on the truth finding problem recently has also been extended to cope with data dynamicity and heterogeneity.

Dynamic data fusion. Jia et al. proposed an incremental strategy adaptive to different update situations by considering the concept drift in learning process [10]. Zhao et al. proposed a probabilistic model that transforms the problem of truth discovery over data streams into a probabilistic inference problem [35]. Hara et al. proposed an incremental data fusion model based on storing provenance information in the form of a sequence of operations by keeping both the original source values and the new fused data in the operations repository [9]. Wang et al. proposed a streaming fact-finding method that recursively updates the previous estimates based on new data [26]. Li et al. investigated the temporal relations among both object truths and source reliability, and proposed an incremental truth discovery framework to dynamically update object truths and source weights [15].

Big data integration. Dong et al. explored the challenges faced by big data integration on the topics of schema mapping, record linkage and data fusion [5]. Sleeman and Finin described a way to subdue VBD that uses popular natural language processing techniques [24]. More recently, Lin and Chen proposed a scheme that integrates domain expertise knowledge to achieve a more precise estimation of source reliability [16], and Yang proposed a graph-based model to conduct truth discovery from conflicting multi-valued objects [6].

All the above methods applied the same basic heuristic idea: a claim is likely to be true if it is provided by trustworthy sources and a source is trustworthy if most its claims are true. Our work is based on our observation that the attribute and object reliabilities may be inconsistent wrt the same source for which no prior work is known. It differs from the existing work in three aspects. First, we use an optimization framework to obtain the fine-grained source reliability estimation on attributes and objects to achieve more accurate truth inference. Second, we use a probability graphical model to compute the object reliability estimation more effectively. Third, we introduce the group reliability to effectively compute the observation deviations of all sources and achieve more effective truth inference. The superiority of our methods in comparison with the state-of-the-art baselines without considering attribute and object reliability inconsistency is demonstrated in the experiment results.

III. Truth Finding by Attribute Reliability Estimation

In this section, we present our TFAR model. The model iteratively updates attribute weights and truths for multi-source data. We formulate the truth finding problem as an optimization problem and obtain its solution of the set of estimated truths and attribute reliabilities by minimizing the weighted deviation summation between the truths and observations. We present several hardness calculation methods and loss functions to complete the attribute weight assignment and truth computation procedure.

A. Basic Definitions

In this part, we will introduce the related concepts and notations, the problem to be solved and solution approach.

Tables 3 and 4 respectively list the related concepts and notations to be used throughout the paper.

Given all the data sources, we aim to find the most trustworthy value for every entity, and infer the reliability degree of each attribute simultaneously. Note that a higher w_{kn} in Table 4 indicates that attribute n is more reliable in source k and observations from this attribute are more likely to be accurate. This is under the observation that if a fact is provided by many trustworthy sources, it is more likely to be true. Furthermore, a source that provides more true facts will be likely to provide more true facts. The source reliability and fact confidence are determined by each other and the true facts are more consistent than false facts, and hence are more likely to be found at the end.

The general approach we use to solve this problem is to calculate first the attribute reliabilities and then the entry truths by iteratively minimizing the deviation (from the truths) summation of all entries weighted by fact-hardness regulated attribute reliabilities. In this approach, we take the collection of observations made by all the sources as the INPUT. The OUTPUT contains an attribute weight list and a truth table. The initial truths are generated by voting and mean methods.

Concept	Explanation
object	a person or thing of interest. e.g., "Question 1".
attribute	an attribute to describe the object. e.g., "Text Analysis".
source	describes the place where information about objects' properties can be collected. e.g., Susan database.
observation	the data describing an attribute of an object from a source. e.g., Text Analysis's Question 1 from Susan database "essay11"
entry	an attribute value of an object. e.g., "Text Analysis's value of Question 1"
truth	accurate information of an entry, which is unique. e.g., the real answer of Text Analysis's Question 1.

Table III: **Summary of terminologies**

Notation	Description
K	Number of sources
N	Number of attributes
M	Number of objects
W	The trustworthiness list of all attributes in all sources$w_{11} \dots w_{1N}, w_{21} \dots w_{2N}, \dots, w_{K1} \dots w_{KN}$
$v_{nm}^{(k)}$	The observation of the n-th attribute for the m -th object made by one in k-th source
$v_{nm}^{(*)}$	The truth for the n-th attribute of the m-th object
d_n	The deviation function for the n-th attribute
w_{kn}	The trustworthiness of the n-th attribute in the k-th source
$S^{(k)}$	The collection of observations made on all the objects by the k-th source $v_{11}^k \dots v_{1M}^k, \dots, v_{N1}^k \dots v_{NM}^k$
S^*	Set of truth for all objects on all properties$v_{11}^* \dots v_{1M}^*, \dots, v_{N1}^* \dots v_{NM}^*$
δ	The threshold of successive truth table entry difference

Table IV: **Summary of notations**

The hardness and deviation calculation methods will be stated later. The iteration procedure will stop if the successive truth table entry difference is below a given threshold δ that will be discussed later.

B. The TFAR Framework

We propose the following optimization framework TFAR that utilizes attribute weight to describe the reliability of sources.

Given M objects, each with N attributes (properties), a set of observations (values) on all the attributes of the objects made from K sources, and an attribute weight (reliability) budget of 1, with the attribute reliability updated periodically, the more reliable an attribute is, the closer the observations on it to the truth is. Thus we should minimize the summation of weighted deviations from the truths to the multi-source observations, where the weights reflect the reliability degrees of the attributes. Summing up, we have the following optimization framework:

$$\min_{S^{(*)},W} f(S^{(*)}, W) = \sum_{k=1}^{K} \sum_{n=1}^{N} \left(w_{kn} * \sum_{m=1}^{M} d(v_{nm}^{(*)}, v_{nm}^{(k)}) \right)$$
$$s.t. \quad \xi(W) = 1 \tag{1}$$

Through minimizing the above function, we will obtain two sets of variables, $S^{(*)}$ representing truths and W representing weights assigned to attributes under the given budget. Loss function d measures the deviation from the observation $v_{nm}^{(k)}$ to the truth $v_{nm}^{(*)}$. It outputs a high value if the deviation is high and low value otherwise. Weight w_{kn} reflects the trustworthiness of the n-th attribute in the k-th source. The higher of w_{kn}, the more trustable of the attribute. Naturally, the truths will rely on the attribute with higher weights to minimize the overall deviations. $\xi(W)$ is the aggregation of the attribute weight assignments under a distribution function. It constrains the weights into a certain range to rationalize the optimization problem.

We iteratively conduct the following three steps to get the attribute weights and the truths through a joint procedure.

First, *entity hardness calculation*. Calculate the hardness of every observation in the truth table by computing the dispersion degree of the corresponding observations in the INPUT sources. We will discuss the dispersion calculation method in Section 3.3.

$$Hardness(v_{nm}) = f(Dispersion(v_{nm}^k)) \tag{2}$$

Second, *attribute weight update*. Compute the attribute weights based on the differences between the given (ground) truths and the observations made on the attributes from the sources, and then update the weights according to the hardness of the corresponding observations:

$$W \leftarrow \arg\min f(W, Hardness(v_{nm})) \tag{3}$$

Third, *truth update*. Compute the truths of each entry to minimize the weighted difference summation between the truths and the entries (observations on an attribute). By computing the truth for every entry, we can obtain the collection of truths $S^{(*)}$.

$$S^{(*)} \leftarrow \arg\min f(W, \{d(v_{nm}^{(*)}, v_{nm}^{(k)})\}) \tag{4}$$

Algorithm 1 Truth estimation Algorithm

Input: Observations made by K sources: $\{S^{(1)}, \ldots, S^{(K)}\}$.
Output: The true value for each object $S^{(*)} = \{v_{nm}^{(*)}\}_{n=1,m=1}^{N,M}$,
and attribute weights $W = (w_{11} \ldots w_{1N}, w_{21} \ldots w_{2N}, \ldots\ldots, w_{K1} \ldots w_{KN})$.

1: Initialize the truths $S^{(*)}$;//using voting and mean methods
2: Calculate hardness of every entry $H = (h_{11} \ldots h_{1M}, h_{21} \ldots h_{2M}, \ldots\ldots, h_{N1} \ldots h_{NM})$ using (2);
3: **repeat**
4: Update attributes weights $W = (w_{11} \ldots w_{1N}, w_{21} \ldots w_{2N}, \ldots\ldots, w_{K1} \ldots w_{KN})$ according to (3) to reflect attributes' reliability based on the estimated truths and the hardness of observations;
5: **for** $k \leftarrow 1$ to K **do**
6: **for** $m \leftarrow 1$ to M **do**
7: **for** $n \leftarrow 1$ to N **do**
8: Compute the truth of the m-th object on the n-th attribute $v_{nm}^{(*)}$ according to (4) based on the current estimation of attribute weights $\{w_{kn}\}$;
9: **end for**
10: **end for**
11: **until** Convergence criterion is satisfied; //the successive truth table entry difference is below the threshold δ
12: **return** $S^{(*)}$ and W.

Implementation of this framework is given in Algorithm 1. We will elaborate the three steps using example functions in the following.

C. Hardness Calculation

Proposition 1. The entity hardness is presented by the dispersion level of the observations. The higher the dispersion level, the harder the entity.

Example 2. *Assume that the answers' selection probabilities are same for one question. If the dispersion level is high, it indicates that the correct rate is low. If most of the students' answers are consistent, it is more likely to indicate that this question is quite easy, though there may be some exceptional cases that popular answers are wrong which is quite rare.*

There are K sources in the INPUT altogether, so there are at most K observations for one entity. Now we present several hardness calculation methods for different data types.

As for categorical data, we add up the occurrence frequency of each term. If the maximum frequency is less than $\lceil K/2 \rceil$, then the dispersion level is high, and this entity will be labeled as hard. Otherwise, the entity will be labeled as easy. As for continuous data, first we divide the values into several numerical intervals, and then we add up the occurrence frequency of each interval. As for text data, we first draw all keywords of each text by deploying a text mining algorithm on the measure of term frequency, and then add up the occurrence frequency of each keyword. As for image data, first we extract features, then build index, at last we search for the features and add up the occurrence frequency of each feature. If the maximum frequency is less than $\lceil K/2 \rceil$, then the dispersion level is high, and the entity will be labeled as hard. Otherwise, the entity will be labeled as easy.

Example 3. *There are three sources in Table 1, so K=3, $\lceil K/2 \rceil$=2. For Digital Analysis attribute, the first entity maximum frequency is 2 (value 8), greater than $\lceil K/2 \rceil$, not hard. The second entity maximum frequency is 3 (value 12), greater than $\lceil K/2 \rceil$, not hard. The third entity maximum frequency is 1, smaller than $\lceil K/2 \rceil$, hard. So there are 1 hard label and 2 easy labels in Digital attribute. Similarly, there are 3 easy labels in Logical attribute, 3 hard labels in Text attribute, and 3 hard labels in Material attribute.*

D. Attribute Weight Assignment

First, we calculate attribute weight assignment. Since attribute weight assignment is similar to source weight assignment, we assume that weight assignment follows exponential distribution and has the following function in the constraint of Equation (1):

$$\xi(\mathrm{W}) = \sum_{k=1}^{K} \sum_{n=1}^{N} \exp(-w_{kn}) \tag{5}$$

Theorem 4. *Suppose the truths are static , the optimization problem (1) with function (5) is convex, and the global optimal solution is given by*

$$w_{kn} = \log \left(\frac{\sum_{k'=1}^{K} \sum_{n'=1}^{N} \sum_{m'=1}^{M} d(v_{n'm'}^{(*)}, v_{n'm'}^{(k')})}{\sum_{m=1}^{M} d(v_{nm}^{(*)}, v_{nm}^{(k)})} \right) \tag{6}$$

Proof: Since the truths are static, (1) has only one set of variables $W = w_{kn}$. We rewrite (1) by replacing w_{kn} with its distribution $t_{kn} = \exp(-w_{kn})$ to prove the convexity of the optimization problem (1): ■

$$\min\ f(t_{kn}) = \sum_{k=1}^{K} \sum_{n=1}^{N} \left(-\log(t_{kn}) \cdot \sum_{m=1}^{M} d(v_{nm}^{(*)}, v_{nm}^{(k)}) \right) \tag{7}$$

$$s.t. \sum_{k=1}^{K} \sum_{n=1}^{N} t_{kn} = 1 \tag{8}$$

The objective function of (7) is a linear combination of negative logarithm functions, and the constraint is linear in t_{kn}, so (7) is convex. Thus, the optimization problem (1) with constraint (8) is convex, and any local optimum is also global optimum [23].

Then we use the Lagrange multipliers to solve (7). The Lagrangian of (7) is as follows:

$$L(\{t_{kn}\}_{k=1,n=1}^{K,N},\lambda)=\sum_{k=1}^{K}\sum_{n=1}^{N}(-\log(t_{kn})*\sum_{m=1}^{M}d(v_{nm}^{(*)},v_{nm}^{(k)}))+\lambda\left(-\sum_{k=1}^{K}\sum_{n=1}^{N}t_{kn}-1\right) \quad (9)$$

where λ is a Lagrange multiplier. We let the partial derivative of Lagrangian with respect to t_{kn} be 0 in order to obtain the optimal value of λ that maximizes the dual objective (9). The solution of the dual presents a feasible solution to the primal problem (7) and (1) according to Lagrangian duality. Hence we have

$$\lambda\sum_{k=1}^{K}\sum_{n=1}^{N}t_{kn}=\sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{m=1}^{M}d(v_{nm}^{(*)},v_{nm}^{(k)}) \quad (10)$$

From the constraint (8), $\sum_{k=1}^{K}\sum_{n=1}^{N}t_{kn}=1$, Equation (10) becomes

$$\lambda=\sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{m=1}^{M}d(v_{nm}^{(*)},v_{nm}^{(k)}) \quad (11)$$

On the other hand, for a fixed k and n, from (10) we have

$$\lambda t_{kn}=\sum_{m=1}^{M}d(v_{nm}^{(*)},v_{nm}^{(k)}) \quad (12)$$

Combining (11) and (12) we have

$$t_{kn}=\frac{\sum_{m=1}^{M}d(v_{nm}^{(*)},v_{nm}^{(k)})}{\sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{m=1}^{M}d(v_{nm}^{(*)},v_{nm}^{(k)})} \quad (13)$$

Because and $w_{kn}=-\log(t_{kn})=\log\frac{1}{t_{kn}}$, we obtain (6).

After we have calculated the deviations of all the entries, we can compute the attribute weights directly using (8).

This weight calculation formula indicates that an attribute with observations closer to the truths will have greater weights. Therefore, (5) is a reasonable constraint function that leads to a meaningful attribute weight assignment formula.

Second, we apply weight regulation. As we stated above, we should update the attribute reliability according to the fact hardness label obtained by Section 2.3 to obtain a more accurate truth table. If we get α hard labels and β easy labels for an $n-$attribute, the attribute reliability can be adjusted as:

$$w_{kn}=w_{kn}*(\frac{M+\alpha}{M+\beta}) \quad (14)$$

Equation (14) shows that the attributes that get answers for harder questions should have a higher reliability relatively. In contrast, the sources that get answers for easier questions should have a lower reliability.

E. Truth Computation

When the attribute weights are fixed, the truth computation is dependent on the data type and loss function. The truth computation methods for categorical data, continuous data, text data, image data and video data are given respectively as follows.

The most commonly used loss function for categorical data is *0-1 loss* in which an error occurs if the observation is different from the truth. Formally, if the n-th attribute is categorical, the deviation from the truth $v_{nm}^{(*)}$ is defined as:

$$d(v_{nm}^{(*)},v_{nm}^{(k)})=\begin{cases}1 & v_{nm}^{(*)}\neq v_{nm}^{(k)},\\ 0 & otherwise.\end{cases} \quad (15)$$

Plugging (15) into the objective function in (1), we can obtain the following formula:

$$v_{nm}^{(*)}\leftarrow\arg\min_{v}\sum_{k=1}^{K}\sum_{n=1}^{N}w_{kn}\cdot d(v,v_{nm}^{(k)}) \quad (16)$$

This formula indicates that based on 0-1 loss function, to minimize the objective function, the truth should be the value that receives the highest weighted votes among all possible values.

Similarly, the L2-normalized loss function for continuous data is (17), indicating that we can use *weighted mean* method to calculate the truth. The truth could be the weighted mean summation of all the observations.

$$d(v_{nm}^{(*)},v_{nm}^{(k)})=\parallel v_{nm}^{(*)}-v_{nm}^{(k)}\parallel_2 \quad (17)$$

For text data, the loss function is (18), indicating that we can use weighted *cosine similarity* method [1] to calculate the deviation.

$$d(v_{nm}^{(*)},v_{nm}^{(k)})=\frac{v_{nm}^{(*)}*v_{nm}^{(k)}}{(\mid v_{nm}^{(*)}\mid*\mid v_{nm}^{(k)}\mid)} \quad (18)$$

For image data, the loss function is (19), indicating that we can use weighted *SITF (Scale Invariant Feature Transform)* [17] method to calculate the deviation.

$$d(v_{nm}^{(*)},v_{nm}^{(k)})=SITF(v_{nm}^{(*)},v_{nm}^{(k)}) \quad (19)$$

For video data, the loss function is (20), indicating that we can use weighted *PSNR (Peak signal-to-noise ratio)* [2] method to calculate the deviation.

$$d(v_{nm}^{(*)},v_{nm}^{(k)})=PSNR(v_{nm}^{(*)},v_{nm}^{(k)}) \quad (20)$$

The above computation follows the principle that an observation stated by reliable sources will be more likely to be regarded as the truth. If the difference between the successive truth table entries is below the threshold δ twice, then the iteration procedure ends. We assume δ is set to be one tenth of the difference.

Example 5. *We calculated the deviations by (15)~(20), the attribute weights by (6). The attribute weights are {(0.45, 0.83, 0.84, 0.52); (0.65, 0.73, 0.84, 0.68); (0.75, 0.69, 0.87, 0.74)} respectively. Given the hardness labels in Example 3 {(3,0); (2,1); (0,3); (3,0) / (α, β)}, by (14) we can get the regulated weights {(0.9, 1.04, 0.42, 1.04); (1.3, 0.91, 0.42, 1.36); (1.5, 0.86, 0.43, 1.48)}. Material analysis is the most reliable attribute in the Leo database as it provides few errors and answers harder questions.*

IV. Truth Finding by Object Reliability Estimation

In this section, we first introduce the measurement of our object reliability. Then we formally describe the proposed truth finding by object reliability estimation model TFOR which is a Bayesian network that incorporates quality measurements into object reliability estimation. We formulate it as a Maximum Likelihood Estimation problem and apply Collapsed Gibbs Sampling to jointly estimate the object reliability and truth. The model iteratively updates the object quality and truth. More details about the model are discussed below.

A. Measurement of Object Reliability

Given the observations and the ground truths of all entities in one object, we can produce the confusion matrix describing their consistency as in a classification method. The definitions of precision, accuracy, recall and specificity are the same. Since 1-sensitivity can capture the false negative rate of conservative objects and specificity can seize the false positive rate of venturous objects, we will use sensitivity and specificity as the measurement of our object reliability. They can cover the quality spectrum in the confusion matrix of an object. In contrast, precision considers only positive claims and accuracy takes only negative claims into consideration, making them unable to recognize both positively and negatively erroneous data because both measurements overlook the fundamental differences of these two types of errors.

Our proposed TFOR model will incorporate specificity and sensitivity into the object estimation procedure. The probability graphical model will create two separate random variables for each object. Other prior knowledge about the object quality can also be easily incorporated into the Bayesian network by specifying the variables. However, without knowing the ground truths, we have to let the object quality and truths as the latent variables infer each other iteratively. We will initialize the truths first, and then calculate the object quality and truths back and forth until meeting the stopping criteria.

B. Model Description

The TFOR model takes object quality, truths of entity and observations as prior knowledge, and models them as random variables. We take specificity and sensitivity as the quality measurement for objects. We model the truth label as a Boolean random variable to help generate the confusion matrix of the objects. As for the observation node, we take the truth label of the entity it refers to and the quality of object it comes from as its father nodes in the graphical model. In this way, we can simulate the four possible real-world outcomes naturally. Then, given the observations, we can infer the truth label and object reliability inversely by Bayes rule. In general, the specificity is high as most objects provide less erroneous data. Similarly, we can take any prior knowledge about these random variables into the model to take advantage of them. Otherwise, we can just use a uniform prior in order to avoid causing bias.

Figure 2 shows the graphical structure of conditional dependence of our proposed TFOR model. In contrast to the existing

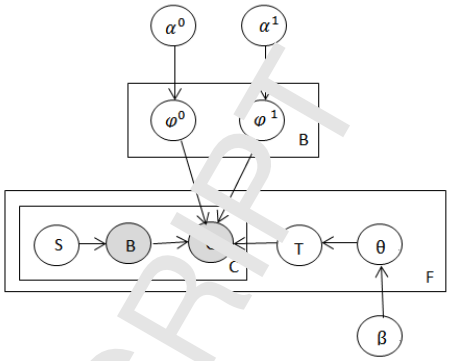


Figure 2: The probabilistic graphical model of TFOR

methods in multi-source aggregation, we jointly model object reliability and truths to simultaneously help each other rather than using source reliability. The generative process of the proposed model is below.

For object reliability:

Draw $\varphi_m^0 \sim Beta(\alpha_{0,1}, \alpha_{0,0})$, where φ_m^0 is the false positive rate of object m, $\alpha_{1,0}$ is the prior false positive count and $\alpha_{0,0}$ is the prior true negative count of each object.

Draw $\varphi_m^1 \sim Beta(\alpha_{1,1}, \alpha_{1,0})$, where φ_m^1 is the sensitivity of object m, $\alpha_{1,1}$ is the prior true positive count and $\alpha_{1,0}$ is the prior false negative count of each object.

For entity truth:

Draw θ_f ~ Beta(β_0,β_1), where θ_f is the prior truth probability of entity f, β_0 is the prior false count and β_1 is the prior true count.

Draw t_f ~ Bernoulli(θ_f), where t_f is the truth label of entity f.

Draw b_m~ Set(S), where $b_m$is the object set in source k, and S is the set of all sources.

Draw o_c~ Bernoulli($\varphi_{s_c}^{t_f}$), where o_c is the observation c of entity f provided by object m.

From the above we can see that α_0 and α_1 control our prior belief of source sensitivity, β expresses the prior belief of how likely each claim is to be true. We will use a uniform prior as it is reasonable to assume the claims are equally likely to be true. We can also see that t_f is a Boolean variable and the prior probability that t_f is true is exactly θ_f.

Remark. Beta distribution is the conjugate of Bernoulli and Binomial distribution, it will make the inference more efficient.

C. Inference Algorithm

In this section, we will introduce a generative process for truth and object reliability estimation of the proposed model and discuss how to infer parameters using Gibbs-EM .

1) Likelihood Functions: Based on the dependencies of the random variables in the TFOR graphical model, the probability of each observation c of entity f given object quality $\varphi_{B_c}^0, \varphi_{B_c}^1$ and truth label θ_f is:

$$p(o_c|\theta_f, \varphi_{B_c}^0, \varphi_{B_c}^1) = p(o_c|\varphi_{B_c}^0)(1-\theta_f) + p(o_c|\varphi_{B_c}^1)\theta_f \quad (21)$$

Then the likelihood of all parameters in the TFOR model given the hyper parameters $\alpha_0, \alpha_1, \beta$ is:

$$p(o,b,t,\theta,\varphi^0,\varphi^1|\alpha_0,\alpha_1,\beta) = \prod_{b\in B} p(\varphi_B^0|\alpha_0)p(\varphi_B^1|\alpha_1)*$$

$$\prod_{f\in F}(p(\theta_f|\beta)\sum_{t_f\in 0,1}\theta_f^{t_f}(1-\theta_f)^{1-t_f}\prod_{c\in C_f}p(o_c|\varphi_{B_c}^{t_f})) \quad (22)$$

2) Estimating Truth: Given the likelihood of all parameters in our model and the observations, the maximum posterior estimation for t is:

$$t_{MAP} = \arg\max_t \int\int\int p(o,b,t,\theta,\varphi^0,\varphi^1)d\theta d\varphi^0 d\varphi^1 \quad (23)$$

If we search the space of all possible truth assignments for $t_{\{MAP\}}$, it would be prohibitively inefficient. So we use Collapsed Gibbs Sampling method to speed up the inference algorithm. Gibbs sampling is used to generate the sequence of samples, whose stationary distribution is what we want to estimate.

Let $t_{-f} = \{f' \in F, f' \neq f\}$. We iteratively sample the truth for each entity given the current truth labels of other entities:

$$p(t_f = i|t_{-f},o,b) \propto \beta_i \prod_{c\in C_f} \frac{n_{B_c,i,o_c+\alpha_{i,o_c}}^{-f}}{n_{B_c,i,0}^{-f}+n_{B_c,i,1}^{-f}+\alpha_{i,0}+\alpha_{i,1}} \quad (24)$$

where $n_{B_c,i,j}^{-f} = |\{c' \in C_{-f}|b_{c'} = b_c, t_{f_{c'}} = i, o_{c'} = j\}|$, i.e., the number of b_c's claims whose observation is j, and referred entity is not f and its truth is i. These counts reflect the quality of b_c based on claims of entities other than f, e.g., $n_{B_c,0,0}^{-f}$ is the number of true negative claims of b_c, $n_{B_c,0,1}^{-f}$ is the false positive count, $n_{B_c,1,0}^{-f}$ is the false negative count, and $n_{B_c,1,1}^{-f}$ is the true positive count.

This procedure implies that the sampling of the truth of each entity is based on the prior for the truths and the object qualities estimated on other entities. We present the pseudo-code of the implementation of the Collapsed Gibbs Sampling in Algorithm 2.

3) Estimating Object Quality: Given the truths estimated in the previous step, we can obtain the predictions of the object quality information from the TFOR model.

Since the posterior of object quality is also a Beta distribution, a maximum posterior estimate of the object quality has a closed-form solution as follows:

$$sensitivity(b) = \varphi_b^1 = \frac{E[n_{b,1,1}]+\alpha_{1,1}}{E[n_{b,1,0}]+E[n_{b,1,1}]+\alpha_{1,0}+\alpha_{1,1}} \quad (25)$$

$$specificity(b) = 1-\varphi_b^0 = \frac{E[n_{b,0,0}]+\alpha_{0,0}}{E[n_{b,0,0}]+E[n_{b,0,1}]+\alpha_{0,0}+\alpha_{0,1}} \quad (26)$$

Algorithm 2 Collapsed Gibbs Sampling for Truth Finding

```
//Initialization step
for all f ∈ F do
    sample t_f from random()
    if (random<ρ) then t_f =1;
    else t_f =0;
    for all c ∈ C_f do
        n_{B_c,t_f,o_c}++;
    //end for
//end for
//Sampling step
for i←1 to K do
    i++;
    for all f ∈ F do
        p_{t_f} ← β_{t_f}, p_{1−t_f} ← β_{1−t_f}
        for all c ∈ C_f do
            p_{t_f} = p_{t_f} * (n_{B_c,t_f,o_c} − 1 + α_{t_f,o_c}) / (n_{B_c,t_f,1} + n_{B_c,t_f,0} − 1 + α_{t_f,1} + α_{t_f,0})
            p_{1−t_f} = p_{1−t_f} * (n_{B_c,1−t_f,o_c} − 1 + α_{1−t_f,o_c}) / (n_{B_c,1−t_f,1} + n_{B_c,1−t_f,0} − 1 + α_{1−t_f,1} + α_{1−t_f,0})
        //Sample t_f from conditional distribution
        if random < p_{1−t_f} / (p_{t_f} + p_{1−t_f}) then
            t_f ← 1 − t_f;
            n_{B_c,1−t_f,o_c}−−; //update the counts
            n_{B_c,t_f,o_c}++; //update the counts
        //calculate the expectation of t_f
        if i > burnin and i%thin = 0 then
            p(t_f = 1) ← p(t_f = 1) + t_f;
    //end for
//end for
```

where $E[n_{b,1,1}] = \sum_{c=C,b_c=b,o_c=j} p(t_{f_c} = j)$ is the expected quality counts of object b which depends on the truth probability of each fact b's claims output by Algorithm 2.

We initialize the truths using the voting and mean method for each entity and calculate the initial counts for each object. Then we update the object reliabilities in each source, truths, object trustworthiness and quality counts accordingly. The iterative procedure will stop after the Gibbs sampling reaches a steady state.

V. Truth Finding by Group Reliability Estimation

In this Section, we will present our truth finding model by group reliability estimation model TFGR. We formulate this model as an optimization problem that updates object groups, group weights and truths iteratively by minimizing the weighted deviation summation between the truths and observations. We also present the methods to obtain the object group weights and truths.

A. Basic Definitions

We first introduce the related concepts and notations used in the TFGR model as well as the problem statement. We use data source 1 (Table 5 (a)) as an example to explain these concepts.

Object	Education	Profession	Phone	Address	City
User1	High school	Builder	9078****35	Add1	Akiachak
User2	Doctor	Student	9078****25	Add5	Akiachak
User3	Master	Engineer	9070****11	Add3	Anchorage
User4	Doctor	Professor	9073****16	Add4	Anchorage

(a) Data source 1

Object	Education	Profession	Phone	Address	City
User1	High school	Sales	9078****35	Add1	Akiachak
User2	Bachelor	Manager	9075****01	Add2	Atmautluak
User3	Master	Manager	9070****11	Add3	Anchorage
User4	Doctor	CEO	9073****16	Add6	Anchorage

(b) Data source 2

Object	Education	Profession	Phone	Address	City
User1	High school	Builder	9078****12	Add1	Akiachak
User2	Doctor	Student	9075****01	Add5	Akiachak
User3	Master	Engineer	9070****11	Add3	Anchorage
User4	Doctor	CEO	9073****16	Add6	Anchorage

(c) Data source 3

Table V: Data source set

Object	Education	Profession	Phone	Address	City
User1	High school	Builder	9078****12	Add1	Akiachak
User2	Doctor	Student	9078****25	Add5	Akiachak
User3	Master	Manager	9070****11	Add3	Anchorage
User4	Doctor	CEO	9073****16	Add6	Anchorage

Table VI: Ground truths

Definition 6. *An* object *is an item of interest, e.g., "User1". An object group is a subset of objects in one source. An attribute is an attribute to describe the object, e.g., "Education". A* source *is the place where information about object's attributes can be collected, e.g., Data source 1. An* observation *is the data describing an attribute of an object from a source, e.g., User1's education from data source 1 is High school. An* entity *is an attribute of an object, e.g., "User1's education".* Truth *is the accurate information of an entry, which is unique, e.g., the real User1's education degree.*

Remark. We follow the assumption that every entity has only one correct value rather than multi-truths.

Definition 7. *$S^{(k)}$ is the collection of observations of all objects on all attributes by the k-th source $\{v^k_{11}, \dots v^k_{1M}, \dots v^k_{N1}, \dots v^k_{NM}\}$. Let $S = \{S^1, S^2, \dots S^k, \dots, S^K\}$ be the set of observations that can be taken as* Input. *Each claim c has the format of v^k_{nm}, where n denotes the attribute number, m denotes the object number, k the source number, and v^k_{nm} the observation on attribute n of object m provided by source k. The* Output $S^{(t)}$ *is a collection of truths for all objects on all attributes in the t-th iteration $\{v^t_{11}, \dots v^t_{1M}, \dots v^t_{N1}, \dots v^t_{NM}\}$, where $v^{(t)}_{nm}$ is the truth on the n-th attribute of the m-th object in the t-th iteration, $w^{(t)}_{km}$ the trustworthiness of the m-th object in the k-th source in t-th iteration and $w^{(t)}_{kg}$ the trustworthiness of the g-th object group in the k-th source in t-th iteration.*

Note that a higher w_{km} in Table 3 indicates that the object m is more reliable than other objects in source k and observations from this object are more likely to be accurate. This is under the basic heuristic idea that a claim is more likely to be true if it is provided by trustworthy sources (especially if by many of them) and a source is trustworthy if most its claims are true.

Definition 8. *Problem definition. Given a source set $\{S^1, S^2, \dots S^k, \dots, S^K\}$ with observation set $\{v^1_{11}, \dots v^k_{1M}, \dots v^k_{N1}, \dots v^K_{NM}\}$ for M objects and Nattributes, we want to learn the object* group list *and object reliabilities $W = \{w_{km}\}^{m=1toM}_{k=1toK}$ for each object in each source and the final truth for each entity v^*_{nm}.*

For true claims that are more consistent than false ones, it is reasonable to believe that we will find the true claims at the end. Apart from the basic heuristic idea, we also believe that different object groups may have different trustworthinesses because of their unique characteristics. Using group reliability will enable us to describe the source trustworthiness more effectively.

Example 9. *As shown in Table 5, data sources 1 and 3 have similar source reliabilities while data source 1 is more accurate in User1 object and data source 3 is more accurate in User4 object. That is to say, in data source 1, object 1 is more reliable than other objects. Using the refined group reliability will enable us to infer truths more accurately.*

Based on this reasoning, the proposed TFGR model calculates the group reliabilities and entry truths by iteratively minimizing the deviation summation between the claims and truths weighted by the object group reliabilities. The initial truths are generated by the voting and mean methods. The iteration procedure terminates when the successive truth table entry difference is below the threshold δ.

B. The TFGR Framework

We propose the following optimization framework to utilize the object group weight that describes the reliability degree of the source. The more reliable an object group is, the closer its observations to the truths are. Thus we should minimize the summation of weighted deviations from the truths to the multi-source observations, where the weights reflect the object group reliabilities. This results in our following optimization framework:

$$\min_{S^{(*)},W} f(S^{(*)}, W) = \sum_{k=1}^{K} \sum_{g=1}^{G_k} w^t_{kg} \sum_{m=1}^{m_{kg}} \sum_{n=1}^{N} d(v^{(t)}_{nm}, v^{(k)}_{nm})$$
$$s.t. \xi(W) = 1 \tag{27}$$

Through minimizing the target function, we will get two sets of variables $S^{(t)}$ and W alternately, where $S^{(t)}$ corresponds to the set of truths in the t-th iteration and W represents the object group weight set. Loss function d_n measures the deviation from the observation $v^{(k)}_{nm}$ to the truth $v^{(t)}_{nm}$ in the

t-th iteration. Usually, it outputs a high value if the deviation is high and low value otherwise. Weight w_{kg}^t reflects the trustworthiness of the g-th object group in the k-th source in the t-th iteration. The higher value of w_{kg}^t, the more trustable the object group. $\xi(W)$ reveals the distributions of object group weights.

We iteratively conduct two steps to get the final object group weight set and the truth set through the following procedure.

Step I: Object Group Division. With the initial truths set S^*, we divide the objects into groups according to the deviation correlation between different objects in one source.

$$\{g_{11}, \ldots\ldots, g_{kG_k}\} \leftarrow partition(d(v_{nm}^{(t)}, v_{nm}^k)) \tag{28}$$

Step II: Object Group Weights Update. For static values of the truths, we compute object group weights based on the variations between the truths in the current iteration and the claims made by the corresponding object:

$$W^t \leftarrow \arg\min f(S^{(t)}) \tag{29}$$

Step III: Truths Update. For fixed weight w_{kg}^t of each object group, we update the truth set by minimizing the weighted variations between the current truth set and the corresponding observations. By computing the truth for every entry, we can obtain the collection of truths of $(t+1)$-th iteration $S^{(t+1)}$.

$$v_{nm}^{(t+1)} \leftarrow \arg\min f(W^t, \{d(v_{nm}^{(t)}, v_{nm}^k)\}) \tag{30}$$

The pseudo code of the TFGR method is given in Algorithm 3. The three steps will be elaborated in the following sections.

Algorithm 3 Truth Finding by Group Reliability Estimation

Input: Observations made by K sources: $\{S^{(1)}, \ldots, S^{(K)}\}$
Output: The true value for each entity $S^{(*)} = \{v_{nm}^{(*)}\}_{n=1,m=1}^{N,M}$ and object weights $W = (w_{11} \ldots w_{1M}, w_{21} \ldots w_{2M}, \ldots\ldots, w_{K1} \ldots w_{KM})$.
Initialize the truths $S^{(1)}$ using voting and mean methods
t=1;//the first iteration
repeat
  Divide every source into object groups $\{g_{11}, g_{12}, \ldots g_{1G_1}, \ldots\ldots g_{K1}, \ldots, g_{KG_K}\}$ with group number $\{m_{11}, m_{12}, \ldots m_{1G_1}, \ldots\ldots m_{K1}, \ldots, m_{KG_K}\}$ using Algorithm 2;
  Update object group weights $W^t = (w_{11}^t \ldots w_{1G_1}^t, \ldots\ldots, w_{K1}^t \ldots w_{KG_K}^t)$ according to (3) to reflect groups' reliability based on the estimated truths;
  for $k \leftarrow 1$ to K **do**
    for $g \leftarrow 1$ to G_K **do**
      for $n \leftarrow 1$ to N **do**
        Compute the truth of the objects in the g-th goup on the n-th attribute $v_{ng}^{(t)}$ according to (4) based on the current estimation of object group weights $\{w_{kg}^t\}$;
      end for
    end for
  t++;
until Convergence criterion is satisfied; //the successive //truth table entry difference is below the threshold δ twice
return $S^{(*)}$ and W.

C. Object Group Division

Dividing objects into groups is motivated by the observation that object reliability is not consistent across the entire dataset. Some objects are more trustable than others because of their characteristics. We propose the following three steps to divide the objects into groups where object's reliability is consistent within each group and different across different groups. This will help us infer the truth more effectively. Furthermore, the following object division computation procedure indicates that the closer of object reliabilities are, the higher probability the objects are in one group, and there are M groups at most and one at least.

Step I: Calculate the reliability of each object.

In this step, we are given the inferred truths and observations of each object on all attributes. First, we compute the (truth, observation)-deviation for each entry on each attribute of each object using loss function d. Then we sum up the deviations of each object to obtain the object derivation. As all objects have the same attributes, we can compare their deviations directly.

Step II: Sort the object deviations in descending order.

Step III: Place all the objects in the same group if their reliability differences are below the threshold Δ. At the end of this step, we will get the object groups and group number of each source.

EXAMPLE 1. *Let us consider the records in Table 1(a). Suppose the inferred truths are the values, and we use 0-1 function as the deviation function for categorical attributes, and square function for continuous attributes. We first compute the overall deviation of each object in this data source. The observations of User1 are "High school","Builder","9078****35","Add1","Akiachak", and the inferred truths are "High school","Builder","9078****12","Add1","Akiachak", so the deviation of User1 is 0+0+1+0=1. Similarly, the deviations of User2, User3 and User4 are 1, 1 and 2 respectively. Second, the descending order of the object deviations are 2, 1, 1 and 1, corresponding to User4, User1, User2 and User3. Third, suppose the threshold is 0.5, then User1, User2 and User3 will be in one group because their deviation differences are below the threshold. Finally we obtain two object groups {User4}, {User1, User2 and User3}.*

D. Object Group Weight Assignment

We use the following regularization function to compute the object group weight assignment in the t-th iteration by constraining the summation of formula $\exp(-w_{kg}^t)$:

$$\xi(\mathrm{W})=\sum_{k=1}^{K}\sum_{g=1}^{G_k}\exp(-w_{kg}^{t}) \tag{31}$$

Theorem 10. *Given the truth set, the optimization problem (27) with convex function (31) in the constraint has the global optimal solution given by*

$$w_{kg}^{t}=\log\left(\frac{\sum_{k'=1}^{K}\sum_{m'=1}^{M}\sum_{n'=1}^{N}d(v_{n'm'}^{(t)},v_{n'm'}^{(k')})}{\sum_{n=1}^{N}d(v_{nm}^{(t)},v_{nm}^{(k)})}\right) \tag{32}$$

Proof: As the truths are static , (27) has only one set of variables W. We assume a variable $\theta_{kg}=\exp(-w_{kg}^{t})$ to prove the convexity of the optimization problem (27). Then (27) can be expressed as follows: ■

$$\min f(\theta_{kg})=\sum_{k=1}^{K}\sum_{g=1}^{G_K}\left(-\log(\theta_{kg})\sum_{m=1}^{m_{kg}}\sum_{n=1}^{N}d(v_{nm}^{(t)},v_{nm}^{(k)})\right) \tag{33}$$

$$s.t.\sum_{k=1}^{K}\sum_{g=1}^{G_K}\theta_{kg}=1 \tag{34}$$

The objective function of (33) is a linear combination of negative logarithm functions, and the constraint is linear in θ_{kg}, so (33) is convex. Thus, the optimization problem (2) with constraint (34) is convex, and any local optimum is also global optimum [23].

Then we apply the Lagrange multipliers to solve (33) as follows:

$$L(\{\theta_{kg}\}_{k=1,g=1}^{K,G_K},\lambda)=\sum_{k=1}^{K}\sum_{g=1}^{G_K}(-\log(\theta_{kg})*$$

$$\sum_{m=1}^{m_{kg}}\sum_{n=1}^{N}d(v_{nm}^{(t)},v_{nm}^{(k)})\Bigg)+\lambda(\sum_{k=1}^{K}\sum_{g=1}^{G_K}\theta_{kg}-1) \tag{35}$$

where λ is a Lagrange multiplier Let the partial derivative of Lagrangian with respect to θ_{kg} be 0 and we can get:

$$\sum_{n=1}^{N}d(v_{nm}^{(t)},v_{nm}^{(k)})=\lambda\theta_{kg} \tag{36}$$

From the constraint that $\sum_{k=1}^{K}\sum_{g=1}^{G_K}\theta_{kg}=1$, we can derive that

$$\lambda=\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{n=1}^{N}d(v_{nm}^{(t)},v_{nm}^{(k)}) \tag{37}$$

Plugging (37) and $w_{kg}=-\log(\theta_{kg})$ into (36), we obtain (32).

Since we have calculated the deviations of all the entries in front, then we can compute the object weights directly using (32).

This object calculation formula indicates that an object with observations closer to the truths will have a greater weight. Therefore, (31) is a reasonable constraint function because it leads to a meaningful object weight assignment formula.

Example 11. *Back to the last example. Given the attribute groups {User2}, {User1, User2 and User3} in Example 5, we use Equation 7 to calculate the group weights. The deviations have been calculated in the object division step, so we can get group weights 1, 2, indicating {User1, User2 and User3} is the most reliable group in data source 1 as it has the least error rate.*

E. Truth Computation

Given the object group weight set, we can use (30) to compute the truth set. The loss function in (30) is determined by the attributes, and different attributes' truth computation may be quite different from each other due to their different characteristics. Li et al. [12] discussed the loss functions about categorical and continuous data in detail. The most commonly used loss function for categorical data is 0-1 loss (Equation (38)) in which an error occurs if the observation is different from the truth. One common loss function for continuous data is normalized squared loss (Equation (39)).

Formally, if the m-th object is categorical, the deviation between the observation $v_{nm}^{(k)}$ and the truth $v_{nm}^{(t)}$ is defined as:

$$d(v_{nm}^{(t)},v_{nm}^{(k)})=\begin{cases}1 & v_{nm}^{(t)}\neq v_{nm}^{(k)},\\ 0 & otherwise.\end{cases} \tag{38}$$

If the m-th object is continuous, the deviation function is defined as:

$$d(v_{nm}^{(t)},v_{nm}^{(k)})=\frac{(v_{nm}^{(t)}-v_{nm}^{(k)})^2}{std(v_{nm}^{(1)},\ldots,v_{nm}^{(K)})} \tag{39}$$

This computation procedure is consistent with the basic heuristic that an observation stated by reliable sources will be more likely to be regarded as the truth. As stated in Algorithm 3, the iteration procedure will stop if the difference between the successive truth set entries is below the threshold δ twice. Then we will get the final weight set and truth set.

VI. Experiments

In Section 5.1, we first introduce the performance measures and baseline methods with their parameter settings, and then describe the three real-world datasets. In Section 5.2.1 we compare the performance of the proposed methods with the baselines. We will also show the convergence rate and run time in Section 5.2.2. The experimental results show that the proposed methods can significantly reduce the error rate compared with the state-of-the-art conflict resolution baseline methods.

A. Experiment Setup

1) Performance Measures: Given the observations of all entities in each object, we need to find out the object reliabilities and entity truths using our proposed methods automatically and compare them with the given ground truths to evaluate the performance of these methods. The TFOR

model runs in a semi-supervised form, the TFAR and TFGR models are implemented in an unsupervised form. We use *Error Rate*, *Distance* and *Cos* as the deviation functions for heterogeneous data types. We use confusion matrix to obtain the specificity and sensitivity of objects. *Error Rate* is the inconsistent proportion between the output and the ground truths of categorical data. *Distance* is the mean of the 2-normalized absolute distance between the output and the ground truths of continuous data. *Cos* is the reverse cosine similarity between the output to the ground truths. Confusion matrix is the consistent and inconsistent counts between the output and ground truths. Specificity of object b is the probability of false facts being claimed as false. Sensitivity of object b is the probability of true facts being claimed as true. For all measures, the lower the value, the better performance of the method.

2) Baselines and parameter settings: We mainly compare our methods with the following methods as they are either classical or state-of-the-art.

- Voting: It is a straightforward way to obtain the truths from a set of observations by taking the value with the maximum count without considering source reliability. This method can only be applied to categorical data.
- CRH [12]: Iteratively calculate the source weights and truths by minimizing the weighted deviation between the truths and observations. It can be applied to heterogeneous data types.
- CATD [11]: Detect the truths from conflicting data with the long-tail phenomenon by considering the source reliability and confidence interval of the estimation. It can be applied to numerical data type.
- TEM [36]: Model the truth existence by incorporating three measures in a graphical model, silent rate, false spoken rate and true spoken rate. This method uses source reliability.
- MTF [29]: It is an integrated Bayesian approach to solve the multi-truth discovery problem by taking source features into account and reformulating the multi-truth-finding problem based on the mappings between sources and values.
- FaitCrowd [18]: Capture various expertise levels on different topics using a probabilistic model. It estimates both topical expertise and true answers simultaneously.

3) Environment: All the experiments are conducted on a windows PC with 4 GB RAM, Intel Core i7 CPU, algorithms are implemented in MATLAB R2013a. All the baselines are under the advised parameter settings to achieve their best performance.

4) Data Description: **The Diabetes Dataset.** This dataset is obtained from Weka-3-7 (data mining software) datasets. This dataset has 8 continuous attributes, one categorical attribute and one text type attribute, 768 objects, and 6912 observations. We generate a dataset consisting of 10 multiple conflicting sources by injecting different kinds of noise into different attributes of ground truth. We take the variation dataset as the input to our approach and baseline methods. We change the data randomly to generate the input data source. A parameter α is used to control the reliability degree of each attribute (a lower α indicates the attribute is altered in a lower chance, we use $\alpha = 0.1$ to 0.5). In this way, we simulated a dataset with attribute reliability in various degrees in all data sources. **The Labor Dataset.** The full name of this dataset is Final settlements in labor negotiations in Canadian industry. This dataset was provided by Stan Matwin from Computer Science Dept of University of Ottawa in Canada. It consists of 57 objects with 16 attributes and was monthly publication. This dataset is also the ground truth of our Labor Data Source. We generate this source by injecting different noises into different objects of the ground truth. A parameter μ is used to control the variation degree of each object by varying its value. In this way, we simulated a data source with various attribute reliabilities in different dataset. **The German Credit Dataset.** This dataset was provided by Professor Dr. Hans Hofmann from Okonometrie University in Germany. It contains 1000 instances and 20 attributes (8 categorical attributes, 7 numerical attributes, 5 text type attributes) from 10 sources. The ground truths are also provided.

B. Experiment Results

1) Estimating Truth and Object Reliability: We evaluate the performances of both our methods and baselines on categorical, continuous and text data types using *Error Rate*, *Distance and Cos* respectively in the TFAR framework. We use Equations (25) and (26) to calculate the sensitivity and specificity of the TFGR framework. Similarly, we use Equation (32) and different deviation functions to obtain the object group reliabilities and truth estimations in the TFOR framework. We summarize the performance of all the methods on Diabetes Dataset in Table 7. We can observe that the proposed TFAR approach achieves better performance than all the baselines. This is because the baseline methods either fail to take entity hardness into consideration or cannot deal with heterogeneous data types with the fine-grained attribute reliability. From the comparison we can see that TFAR can model source reliability more accurately by inferring attribute reliability and adjusting the reliability by entity hardness. This also justifies our assumption that attribute reliability is more accurate than sources reliability. The TGOR and TFGR models are not good at modeling the attribute reliabilities.

The quality estimation with truth threshold 0.5 of all methods on Labor Data Set is summarized in Table 8. It is obvious that TFOR performs better than other methods. Because other methods all use source reliability to measure the trustworthiness of a source and ignore the fact that different labor record reliabilities may be different from each other. On the contrary, the proposed TFOR method takes every object in each source as an independent "source" , and computes an object's reliability through inferring the graphical model. The TFGR does not perform as well as TFOR due to its hardness in setting an appropriate difference threshold to bring objects into groups with random object reliabilities. The TFAR model does not perform well in dealing with different object trustworthinesses.

We can observe that the TFGR method performs best in quality estimation on the German Credit Dataset in Table 9.

Method	Specificity	Sensitivity
TFAR	0.97	0.95
TFOR	0.86	0.84
TFGR	0.85	0.82
CRH	0.9	0.89
MTF	0.81	0.78
CATD	0.78	0.75
Voting	0.65	0.72

Table VII: Performance Comparison on Diabetes Dataset

Method	Specificity	Sensitivity
TFAR	0.81	0.85
TFOR	1.0	0.96
TFGR	0.98	0.91
TEM	0.91	0.85
MTF	0.75	0.88
FaitCrowd	0.90	0.90
Voting	0.64	0.74

Table VIII: Performance Comparison on Labor Data set

The TFGR method divides several object groups according to their reliabilities. It coincidences with the common sense that there exist several credit levels in a Credit System as different levels enjoying quite different rights.

Through the above experiments we can draw the conclusion that the TFAR model performs well when the attribute trustworthiness are different from each other, and the TFOR model performs well when the object reliabilities are inconsistent with each other while the TFGR model performs better when there are several object reliability levels among all objects.

2) Efficiency: **Convergence rate.** Since our inference algorithm is an iterative procedure, we now show the convergence rate using Credit Dataset. We make 4 sequential predictions using the samples in the first 10 iterations with sample gaps 0, 1, 2, 3 respectively. We repeat 5 runs to count for randomization due to sampling and compute the average specificity and sensitivity which are shown in Figure 3. We can see that they can reach stable after only 5 iterations, showing that the proposed method converges quickly in practice. **Runtime.** All three methods have a linear time complexity on the number of claims in the data. To achieve the same accuracy, TFGR takes less time than TFOR as it computes the deviations more efficiently in groups rather than individual objects.

Method	Specificity	Sensitivity
TFAR	0.94	0.88
TFOR	0.99	0.97
TFGR	1.0	0.98
TEM	0.94	0.81
MTF	0.67	0.83
FaitCrowd	0.85	0.90
Voting	0.62	0.78

Table IX: Performance Comparison on German Credit Dataset

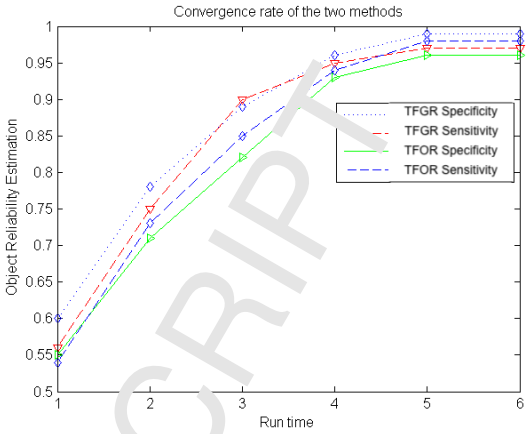


Figure 3: Convergence Rate

VII. Conclusion

We proposed three effective models, TFAR, TFOR and TFGR, for truth finding by estimating the reliabilities on heterogeneous attributes, objects and object groups in a multi-source environment respectively. In our TFAR model, we gave an optimization framework and its implementation algorithm for computing attribute reliabilities to achieve more accurate description of source trustworthiness. In our TFOR model, we presented a generative process to obtain the object reliability by regarding truth as a latent variable and applying the Bayesian approach that can incorporate prior knowledge about the truths of objects, and developed an efficient inference algorithm based on Gibbs sampling to infer the truths. In our TFGR model, we presented an optimization framework and its implementation algorithm for iteratively computing object group reliabilities. Experiments on three real world datasets show that the proposed methods have better performance than the classical and state-of-the-art baseline methods.

There are still interesting challenges on this problem. Our method is based on the intuition that attributes are independent with each other and their values are static. However, these assumptions may not always hold (e.g., a person's title may have relationship with his age, an attribute's value may change over time). As our future work, we will extend our study to take into consideration of the relationship among attributes and dynamic changes of attribute values in order to gain deep insight into formulated problem and the behavior of its solution. We will also incorporate prior knowledge of object characteristics such as truth counts etc. into the TFOR mode to obtain a more accurate description of the object reliability.

Acknowledgement

This work is supported by National Key R & D Program of China Project #2017YFB0203201 and Australian Research Council Discovery Projects funding DP150104871. The corresponding author is Hong Shen.

References

[1] Courtney Corley and Rada Mihalcea. Measuring the semantic similarity of texts. *Unt Scholarly Works*, pages 13–18, 2005.

[2] Johannes F De Boer, Barry Cense, B Hyle Park, Mark C Pierce, Guillermo J Tearney, and Brett E Bouma. Improved signal-to-noise ratio in spectral-domain compared with time-domain optical coherence tomography. *Optics letters*, 28(21):2067–2069, 2003.

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

[4] Xin Luna Dong, Barna Saha, and Divesh Srivastava. Less is more: Selecting sources wisely for integration. *Proceedings of the VLDB Endowment*, 6(2):37–48, 2012.

[5] Xin Luna Dong and Divesh Srivastava. Big data integration. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, pages 1245–1248. IEEE, 2013.

[6] Xiu Susie Fang. Truth discovery from conflicting multi-valued objects. In *WWW (Companion Volume)*, pages 711–715. ACM, 2017.

[7] Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. Corroborating information from disagreeing views. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 131–140. ACM, 2010.

[8] Liang Ge, Jing Gao, Xiaoyi Li, and Aidong Zhang. Multi-source deep learning for information trustworthiness estimation. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 766–774. ACM, 2013.

[9] Carmem Satie Hara, Cristina Dutra de Aguiar Ciferri, and Ricardo Rodrigues Ciferri. Incremental data fusion based on provenance information. In *In Search of Elegance in the Theory and Practice of Computation*, pages 339–365. Springer, 2013.

[10] Li Jia, Hongzhi Wang, Jianzhong Li, and Hong Gao. Incremental truth discovery for information from multiple data sources. In *Web-Age Information Management*, pages 56–66. Springer, 2013.

[11] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. A confidence-aware approach for truth discovery on long-tail data. *Proceedings of the VLDB Endowment*, 8(4):425–436, 2014.

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

[13] Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. Truth finding on the deep web: is the problem solved? In *Proceedings of the 39th international conference on Very Large Data Bases*, pages 97–108. VLDB Endowment, 2012.

[14] Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. A survey on truth discovery. *arXiv preprint arXiv:1505.02463*, 2015.

[15] Yaliang Li, Qi Li, Jing Gao, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. On the discovery of evolving truth. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 675–684. ACM, 2015.

[16] Xueling Lin and Lei Chen. Domain-aware multi-truth discovery from conflicting sources. *PVLDB*, 11(5):635–647, 2018.

[17] Tony Lindeberg. Scale invariant feature transform. *Scholarpedia*, 7(5):10491, 2012.

[18] Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 745–754. ACM, 2015.

[19] Adway Mitra and Srujana Merugu. Reconciliation of categorical opinions from multiple sources. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1561–1564. ACM, 2013.

[20] Aditya Pal, Vibhor Rastogi, Ashwin Machanavajjhala, and Philip Bohannon. Information integration over time in unreliable and uncertain environments. In *Proceedings of the 21st international conference on World Wide Web*, pages 789–798. ACM, 2012.

[21] Ravali Pochampally, Anish Das Sarma, Xin Luna Dong, Alexandra Meliou, and Divesh Srivastava. Fusing data with correlations. Sigmod, 2014.

[22] Theodoros Rekatsinas, Xin Luna Dong, and Divesh Srivastava. Characterizing and selecting fresh data sources. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 919–930. ACM, 2014.

[23] S.Boyd and L.Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[24] Jennifer Sleeman and Tim Finin. Taming wild big data. In *AAAI Fall Symposium on Natural Language Access to Big Data*. AAAI Press, 2014.

[25] Dalia Attia Waguih and Laure Berti-Equille. Truth discovery algorithms: An experimental evaluation. *arXiv preprint arXiv:1409.6428*, 2014.

[26] Dong Wang, Tarek Abdelzaher, Lance Kaplan, and Charu C Aggarwal. Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications. In *Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference on*, pages 530–539. IEEE, 2013.

[27] Dong Wang, Lance Kaplan, Hieu Le, and Tarek Abdelzaher. On truth discovery in social sensing: A maximum likelihood estimation approach. In *Proceedings of the 11th international conference on Information Processing in Sensor Networks*, pages 233–244. ACM, 2012.

[28] Xianzhi Wang, Quan Z Sheng, Xiu Susie Fang, Xue Li, Xiaofei Xu, and Lina Yao. Approximate truth discovery via problem scale reduction. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 503–512. ACM, 2015.

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

[30] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, pages 2035–2043, 2009.

[31] Xiaoxin Yin, Jiawei Han, and Philip S Yu. Truth discovery with multiple conflicting information providers on the web. *Knowledge and Data Engineering, IEEE Transactions on*, 20(6):796–808, 2008.

[32] Xiaoxin Yin and Wenzhao Tan. Semi-supervised truth discovery. In *Proceedings of the 20th international conference on World wide web*, pages 217–226. ACM, 2011.

[33] Bo Zhao and Jiawei Han. A probabilistic model for estimating real-valued truth from conflicting sources. In *Proc. of 10th Intl. Workshop on Quality in Databases, in conjunction with VLDB*, pages 1–7, 2012.

[34] Bo Zhao, Benjamin IP Rubinstein, Jim Gemmell, and Jiawei Han. A bayesian approach to discovering truth from conflicting sources for data integration. *Proceedings of the VLDB Endowment*, 5(6):550–561, 2012.

[35] Zhou Zhao, James Cheng, and Wilfred Ng. Truth discovery in data streams: A single-pass probabilistic approach. 2014.

[36] Shi Zhi, Bo Zhao, Wenzhu Tong, Jing Gao, Dian Yu, Heng Ji, and Jiawei Han. Modeling truth existence in truth discovery. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1543–1552. ACM, 2015.