A Confidence-Aware Approach for Truth Discovery on Long-Tail Data

Qi Li[1], Yaliang Li[1], Jing Gao[1], Lu Su[1],
Bo Zhao[2], Murat Demirbas[1], Wei Fan[3], and Jiawei Han[4]
[1]SUNY Buffalo, Buffalo, NY USA
[2]Microsoft Research, Mountain View, CA USA
[3]Huawei Noah's Ark Lab, Hong Kong
[4]University of Illinois, Urbana, IL USA
{qli22,yaliangl,jing,lusu}@buffalo.edu, bozha@microsoft.com,
demirbas@buffalo.edu, david.fanwei@huawei.com, hanj@illinois.edu

ABSTRACT

In many real world applications, the same item may be described by multiple sources. As a consequence, conflicts among these sources are inevitable, which leads to an important task: how to identify which piece of information is trustworthy, i.e., the truth discovery task. Intuitively, if the piece of information is from a reliable source, then it is more trustworthy, and the source that provides trustworthy information is more reliable. Based on this principle, truth discovery approaches have been proposed to infer source reliability degrees and the most trustworthy information (i.e., the truth) simultaneously. However, existing approaches overlook the ubiquitous long-tail phenomenon in the tasks, i.e., most sources only provide a few claims and only a few sources make plenty of claims, which causes the source reliability estimation for small sources to be unreasonable. To tackle this challenge, we propose a confidence-aware truth discovery (CATD) method to automatically detect truths from conflicting data with long-tail phenomenon. The proposed method not only estimates source reliability, but also considers the confidence interval of the estimation, so that it can effectively reflect real source reliability for sources with various levels of participation. Experiments on four real world tasks as well as simulated multi-source long-tail datasets demonstrate that the proposed method outperforms existing state-of-the-art truth discovery approaches by successful discounting the effect of small sources.

1. INTRODUCTION

Big data leads to big challenges, not only in the *volume* of data but also in its *variety* and *veracity*. In many real applications, multiple descriptions often exist about the same set of objects or events from different sources. For example, customer information can be found from multiple databases in a company, and a patient's medical records may be scattered across different hospitals. Unavoidably, data or information inconsistency arises from multiple sources. Then, among conflicting pieces of data or information, which one is more trustworthy, or represents the true fact? Facing the daunting scale of data, it is unrealistic to expect a human to "label" or tell which data source is reliable or which piece of information is accurate. Therefore, an important task is to automatically infer data trustworthiness from multi-source data to resolve conflicts and find the most trustworthy piece of information.

Proceedings of the VLDB Endowment, Vol. 8, No. 4

Finding Trustworthy Information. One simple approach for this task is to assume that "majority" represents the "truth". In other words, we take the value claimed by the majority sources or take the average of the continuous values reported by sources, and regard it as the most trustworthy fact. The drawback of this simple approach is its inability to characterize the reliability levels of sources. It regards all sources as equally reliable and does not distinguish them, and thus may fail in scenarios when there exist sources sending low quality information, such as faulty sensors that keep emanating erroneous information, and spam users who propagate false information on the Web. To overcome this limitation, techniques have been proposed to simultaneously derive trustworthy facts and estimate source reliability degrees [8, 9, 16, 18–20, 23–26, 31, 34–37]. A common principle to the techniques is as follows. The sources which provide trustworthy information are more reliable, and the information from reliable sources is more trustworthy. In these approaches, the most trustworthy fact, i.e., the *truth*, is computed as a weighted voting or averaging among sources where more reliable ones have higher weights. Although different formulas have been proposed to derive source weights (i.e., reliability degrees), the same principle applies: The source weight should be proportional to the probability of the source giving trustworthy information. In practice, this probability is simulated as the percentage of correct claims of the source. The more claims a source makes, the more likely that this estimation of source reliability is closer to the true reliability degree.

Long-tail Phenomenon. However, sources with very few claims are common in applications. The number of claims made by sources typically exhibits long-tail phenomenon, that is, most of the sources only provide information about one or two items, and there are only a few sources that make lots of claims. For example, although there are numerous websites containing information about one or several celebrities, there are few websites which, like Wikipedia, provide extensive coverage for thousands of celebrities. Another example concerns user participation in survey, review or other activities. On average, participants only show interests to a few items whereas very few participants cover most of the items. Long-tail phenom-

ena are ubiquitous in real world applications, which bring obstacles to the task of information trustworthiness estimation.

Recall that identifying reliable sources is the key to find trustworthy information, and source reliability is typically estimated by the empirical probability of making correct claims. The effectiveness of this estimation is heavily affected by the total number of claims made by each source. When a source makes a large number of claims, it is likely that we can obtain a relatively accurate estimate of source reliability. However, sources with a few claims occupy the majority when long-tail phenomenon exists. For such "small" sources, there is no easy way to evaluate their reliability degrees. Consider an extreme case when most sources only make one claim to one single item. If the claim is correct, its accuracy is one and the source is considered as highly reliable. If the claim is wrong, its accuracy is zero and the source is regarded as highly unreliable. In some sense such an estimate based on one single claim is totally random. When weighted voting is conducted based on the estimates of source reliability, the "unreliable" estimation of source reliability for many "small" sources will inevitably impair the ability of detecting trustworthy information. We illustrate this phenomenon and its effect on the task of truth discovery with more details in Sections 2 and 4.

Limitation of Traditional Approaches. One may argue that one way to tackle the issue of insufficient data for accurate reliability estimation is to remove sources that provide only a few claims. However, this simple strategy suffers from the following challenges. First, we need a threshold on the number of claims to classify "small" or "large" for sources. Second, as the majority of sources claims very few facts, the removal of these sources may result in sparse data and limited coverage. An alternative strategy could be drawn from Bayesian estimation in which a smoothing prior can be added [36, 37]. We can add a fixed "pseudo" count in the computation of source accuracy so that the estimation can be smoothed for sources with very few claims. When there are many sources, typically a uniform prior is adopted, i.e., the same "pseudo" count applies to all sources. How to select an appropriate pseudo count is an open question. Moreover, a uniform prior may not fit all the scenarios, but setting a non-uniform prior is difficult when there is a large number of sources.

Summary of Proposed Approach. In this paper, we propose a confidence-aware approach to detect trustworthy information from conflicting claims, where the long-tail phenomenon is observed in data. We propose that source reliability degree is reflected in the variance of the difference between the true fact and the source input. The basic principle is that an unreliable source will make errors frequently and have a wide spectrum of errors in distribution. To resolve conflicts and detect the most trustworthy piece of information, we take a weighted combination of source input in which the weight of each source corresponds to its variance. Since variance is unknown, we derive an effective estimator based on the confidence interval of the variance. The chi-squared distribution in the estimator incorporates the effect of sample size. The overall goal is to minimize the weighted sum of the variances to obtain a reasonable estimate of the source reliability degrees. By optimizing the source weights, we can assign high weights to reliable sources and low weights to unreliable sources when the sources have sufficient claims. When a source only provides very few claims, the weight is mostly dominated by the chi-squared probability value so that the source reliability degree is automatically smoothed and small sources will not affect the trustworthiness estimation heavily. We apply the proposed method and various baseline methods on four real world application scenarios and simulated datasets. Existing approaches, which regard small and big sources the same way, fail to provide an accurate estimate of truths. In contrast, the proposed method can successfully detect trustworthy information by effectively estimating source reliability degrees.

In summary, we make the following contributions in this paper:

- We identify the pitfalls and challenges in data with long-tail phenomenon for the task of *truth discovery*, i.e., detecting the most trustworthy facts from multiple sources of conflicting information.
- We propose to combine multi-source data in a weighted aggregation framework and search for the best assignment of source weights by solving an optimization problem.
- An estimator based on the confidence interval of source reliability is derived. This estimator can successfully estimate source reliability, and discount the effect of small sources without the hassle of setting pseudo counts or priors.
- We test the proposed algorithm on real world long-tail datasets, and the results clearly demonstrate the advantages of the approach in finding the true facts and identifying reliable sources. We also provide insights about the method by illustrating its behavior under various conditions using simulations.

In the following section, we first describe some real world applications and the collected datasets to illustrate the challenge of long-tail phenomenon in truth discovery tasks. Then, in Section 3, we formulate the problem and derive the proposed method. In Section 4, various experiments are conducted on both real world and simulated datasets, and we validate the effectiveness and efficiency of the proposed method. Related work is discussed in Section 5, and finally, we conclude the paper in Section 6.

2. APPLICATIONS AND OBSERVATIONS

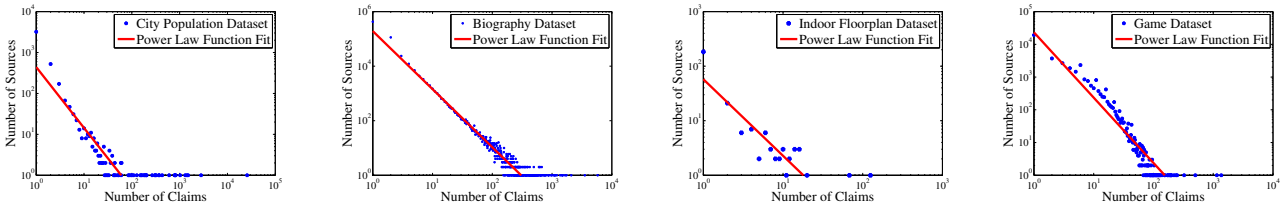
In this section, we present a broad spectrum of real world truth discovery applications where the long-tail phenomenon can be observed. Although the long-tail phenomenon is not rare in truth discovery tasks, it does not receive enough attention yet.

Web Information Aggregation. When the Web becomes one of the most important information origins for most people, it is crucial to analyze the reliability of various data sources on the Web in order to obtain trustworthy information. The long-tail phenomenon is common on the Web. Only a few famous big data sources, such as Wikipedia, may offer plenty of information, but most websites may only provide limited information.

We introduce two specific truth discovery scenarios for web information aggregation: truth discovery on city population and on biography information. For these tasks, we are interested in aggregating the population information about some cities at different years and people's biography respectively. Two datasets[1] were crawled by the authors in [23]. The information can be found in the cities' or persons' Wikipedia infoboxes, and the edit histories of these infoboxes are examined. As Wikipedia pages can be edited by any user, for a specific entity, multiple users may contribute to it. The information from these users is not consistent, and some users may provide more reliable information than the others.

Social Sensing. Social sensing is a newly emerged sensing scenario where the collection of sensory data are carried out by a large group of users via sensor-rich mobile devices such as smartphones. In social sensing applications, human-carried sensors are the sources of information. For the same object or event, different sensors may report differently due to many factors, such as the

[1] http://cogcomp.cs.illinois.edu/page/resource_view/16



(a) City Population Dataset (b) Biography Dataset (c) Indoor Floorplan Dataset (d) Game Dataset

Figure 1: Long-tail phenomenon is observed with real world datasets.

quality of the sensors and the way in which the sensor carrier performs the sensing task. Truth discovery techniques can be useful for social sensing to improve the quality of sensor data integration by inferring the sources' quality. In many social sensing applications, only a few sensors are incessantly active while most of others are activated occasionally, which causes the long-tail phenomenon.

A representative example of social sensing is the construction of indoor floorplans [1, 28]. This research topic has recently drawn a growing interest since it potentially can support a wide range of location-based applications. The goal is to develop an automatic floorplan construction system that can infer the information about the building floorplan from the movement traces of a group of smartphone users. The movement traces of each user can be derived from the readings of inertial sensors (e.g., accelerometer, gyroscope, and compass) built in the smartphone. Here we are interested in one specific task of floorplan construction, i.e., to estimate the distance between two indoor points (e.g., a hallway segment). We develop an Android App that can estimate the walking distances of a smartphone user through multiplying his/her step size by step count inferred using the in-phone accelerometer. When App users are walking along the hallways, we record the distances they have traveled. For the same hallway segment, the estimated distances given by different users are inevitably different due to the varieties in their walking patterns, the ways of carrying the phones, and the quality of in-phone sensors.

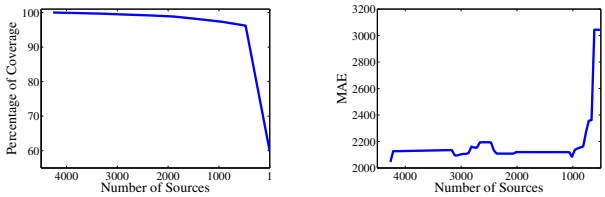
Crowd Wisdom. The wisdom of the crowd can be achieved by integrating the crowd's answers and opinions towards a set of questions. By carefully estimating each participants' abilities, the aggregation among the crowd's inputs can often achieve better answers compared with the answers given by a single expert. Current technologies enable convenient crowd wisdom implementation, and truth discovery provides an effective way to aggregate participants' input and output accurate answers. The long-tail phenomenon happens in the crowd wisdom applications because many participants only show interests in a couple of questions, while a few participants answer lots of the questions.

In this application, we design an Android App as a crowd wisdom platform based on a popular TV game show "Who Wants to Be a Millionaire" [2]. When the game show is on live, the Android App sends each question and four corresponding candidate answers to users, and then collects their answers. For each question, answers from different users are available, and usually these answers have conflicts among them. We can then create a superplayer that outperforms all the participants by integrating answers from all of them.

Due to page limit, we only introduce three applications, but there are more than we can list. In these applications, we observe the difference in information quality of various sources which motivates truth discovery research. Long-tail phenomenon is ubiquitous in these truth discovery tasks. In the following, we demonstrate the long-tail phenomenon using the four truth discovery datasets we experiment on. The four datasets are introduced in the above discussions and more information can be found in Section 4. Their statistical information is summarized in Table 1. We count the number of claims made by each source and Figure 1 shows the distribution of this statistic. The figures witness a clear long-tail phenomenon: Most sources provide few claims and only a small proportion of sources provide a large number of claims. In order to demonstrate the long-tail phenomenon clearer, we further fit the City Population, Biography, Indoor Floorplan and Game datasets into power law function, a typical long-tail distribution[2]. Figure 1 shows that the fitting curves closely match the data, which is a strong evidence of long-tail phenomenon.

Table 1: Statistics of real world long-tail datasets

	City Population	Biography	Indoor Floorplan	Game
# Sources	4107	607819	247	38196
# Entities	43071	9924	129	2169
# Claims	50561	1372066	740	221653



(a) Coverage (b) Performance

Figure 2: The percentage of coverage decreases and MAE increases as more sources are removed.

As discussed in Section 1, removing the sources that provide few claims might be a possible solution. The main shortcoming of this solution is that a large proportion of the whole dataset is discarded. Figure 2 demonstrates two consequences caused by this problem using City Population dataset. All the sources are ordered based on the number of claims they provide. At the very beginning, we consider all sources, and then gradually remove sources starting from the smallest ones. One consequence is the sacrifice of coverage (Figure 2a). If we regard "small" sources as those whose claims are less than 1% of the number of claims made by the biggest source and remove them, the percentage of coverage is 88.07%. In addition to the low percentage of coverage, we lose 10491 claims counting for 20.74% of all claims, which leads to another consequence: performance degrading. Figure 2b shows that the mean absolute error (MAE) increases as more sources are removed (detail of the measure is introduced in Section 4.1). After removing

[2]Note that we use power law distribution as an example of long-tail phenomenon, but long-tail phenomenon is a general scenario and some other distributions, such as Burr distribution and log-normal distribution, can be used to describe long-tail phenomenon too.

the small sources, the number of claims for each entity will shrink dramatically, which causes the problem that the information is not sufficient to estimate trustworthy output.

Smoothing prior or "pseudo" count, as mentioned in Section 1, is another possible solution. The difficulty of this solution lies in setting the "pseudo" count. As Figure 1 illustrates, the numbers of claims made by sources are significantly different. It is unfair to use the same "pseudo" count for all sources. However, with thousands or even hundreds of thousands sources, to assign individual "pseudo" count to each source is unrealistic and impossible to tune.

3. METHODOLOGY

In this section, we describe the proposed method, which tackles the challenge that most of the sources only provide information about few items. We model the truths as weighted combination of the claims from multiple sources and formulate the weight computation as an optimization problem. Some practical issues are discussed at the end of this section.

3.1 Problem Formulation

We start by introducing terminologies and notations used in this paper with an example. Then the problem is formally formulated.

DEFINITION 1. *An* entity *is an item of interest. A* claim *is a piece of information provided by a source about a given entity. A* truth *is the most trustworthy piece of information for an entity.*

DEFINITION 2. *Let $\mathcal{C} = \{c_1, c_2, \dots c_C\}$ be the set of claims that can be taken as input. Each claim c has the format of (n, s, x_n^s), where n denotes the entity, s denotes the source, and x_n^s denotes the information of entity n provided by source s.*

DEFINITION 3. *The output $\mathcal{X}$ is a collection of (n, x_n^*) pairs, where x_n^* denotes the truth for entity n.*

Table 2: A sample census database

Entity	Source ID	Population (million)
NYC	Source A	8.405
NYC	Source B	8.837
NYC	Source C	8.4
NYC	Source D	13.175
DC	Source A	0.646
DC	Source B	0.6
LA	Source A	3.904
LA	Source B	15.904
...	...	...

Table 3: $\mathcal{X}$ and the ground truths for the sample census database

$\mathcal{X}$		Ground truths	
Entity	Population	Entity	Population
NYC	8.423	NYC	8.420
DC	0.645	DC	0.646
LA	4.291	LA	4
...	...	...	...

EXAMPLE 1. *Table 2 shows a sample census database. In this particular example, an entity is a city and a claim is a tuple in the database. Source A states that New York City has a population of 8.405 million, so its corresponding $x_n^s = 8.405$. Note that in this example, x_n^s is a numerical value, but we do not limit x_n^s to be continuous data type only. Discussions on categorical values can be found in Section 3.2.4. Table 3 shows the output $\mathcal{X}$ using the proposed method and the ground truth for this sample census database. Comparing with the ground truths, every source may make some mistakes on their claims, but some sources make fewer errors than the others. For example, source A's claims are closer to the ground truths than source B's claims, which means the former is more reliable than the latter, so source A deserves a higher weight when inferring the truth. Source C seems to be reliable, but based on one claim, it is hard to judge. The proposed method achieves very close results comparing with the ground truths by accurately estimating the source reliability degrees.*

Given input $\mathcal{C}$, our task is to resolve the conflicts and find the most trustworthy piece of information from various sources for every entity. In addition to the truth $\mathcal{X}$, we also simultaneously infer the reliability degree of each source w_s based on input information. A higher w_s indicates that the s-th source is more reliable and information from this source is more trustworthy.

Table 4 summarizes all the notations used in this paper. σ_s^2 and u_s^2 will be introduced in the next subsection.

Table 4: Notations

Notation	Definition
$\mathcal{C}$	set of claims (input)
$\mathcal{N}$	set of entities
n	the n-th entity
$\mathcal{S}$	set of sources
s	the s-th source
N_s	the set of entities provided by source s
S_n	the set of sources that provide a claim on entity n
x_n^s	information for entity n provided by source s
$\mathcal{X}$	set of truths (output)
x_n^*	the truth for entity n
w_s	weight for source s
σ_s^2	error variance of source s
u_s^2	upper bound of variance σ_s^2

3.2 CATD Method

In this section, we formally introduce the proposed method, called *C*onfidence-*A*ware *T*ruth *D*iscovery (CATD), for resolving the conflicts and finding the truths among various sources. The proposed method can handle the challenge brought by the long-tail phenomenon that we observe.

3.2.1 Truth Calculation

Here we only consider the single truth scenario, i.e., there is only one truth for each entity although sources may provide different claims on the same entity.

The basic idea is that reliable sources provide trustworthy information, so the truth should be close to the claims from reliable sources. Many truth discovery methods [8, 16, 19, 20, 23, 23–26, 31, 34–37] use weighted voting or averaging more or less to achieve the truths, which overcome the issue of conventional voting or averaging schema that assumes all the sources are equally reliable.

We propose to use the same weighted averaging strategy to obtain the truths. Since sources are usually consistent in the quality of its claims, we can use source weight, i.e., the source reliability degree, w_s as the weight for all the claims provided by s:

$$x_n^* = \frac{\sum_{s \in S_n} w_s \cdot x_n^s}{\sum_{s \in S_n} w_s}. \quad (1)$$

However, the source reliability degrees are usually unknown *a priori*. Therefore, the key question we want to explore next is how to find the "best" assignment of w_s.

3.2.2 Source Weight Calculation

In this paper, we assume that all sources make their claims independently, i.e., they do not copy from each other. We leave the case when source dependence happens for future work. We can regard that each source's information is independently sampled from a hidden distribution. Errors, which are differences between the claims and the truths, may occur for every source. The variance of the error distribution reflects the reliability degree of this source: if a source is unreliable, the errors it makes occur frequently and have a wide spectrum in general, so the variance of the error distribution is big. We believe that none of the sources make errors on purpose, so the mean of the error distribution, which indicates its bias, is 0. We propose to use Gaussian distribution to describe errors, which is widely adopted in many fields. For each source, its error follows a Gaussian distribution with mean 0 and variance σ_s^2, i.e.,

$$\epsilon_s \sim N(0, \sigma_s^2).$$

Since we have the source independence assumption, the errors that sources make are independent too. We can then compute the distribution for the error of the weighted combination in Eq.(1) as:

$$\epsilon_{combine} \sim N\left(0, \frac{\sum_{s\in\mathcal{S}} w_s^2\sigma_s^2}{(\sum_{s\in\mathcal{S}} w_s)^2}\right), \tag{2}$$

where $\epsilon_{combine} = \frac{\sum_{s\in\mathcal{S}} w_s\epsilon_s}{\sum_{s\in\mathcal{S}} w_s}$. Without loss of generality, we constrain $\sum_{s\in\mathcal{S}} w_s = 1$.

For a Gaussian distribution, the variance determines the shape of the distribution. If the variance is small, then the distribution has a sharp and high central peak at the mean, which indicates a high probability that errors are close to 0. Therefore, we want the variance of the $\epsilon_{combine}$ to be as small as possible. We formulate this goal into the following optimization problem:

$$\begin{aligned} \min_{\{w_s\}} \quad & \sum_{s\in\mathcal{S}} w_s^2\sigma_s^2 \\ \text{s.t.} \quad & \sum_{s\in\mathcal{S}} w_s = 1, w_s \geqslant 0, \forall s \in \mathcal{S}. \end{aligned} \tag{3}$$

Usually the theoretical σ_s^2 is unknown for each source. Inspired by sample variance, the following estimator can be used to estimate the real variance σ_s^2:

$$\hat{\sigma_s^2} = \frac{1}{|N_s|}\sum_{n\in N_s}\left(x_n^s - x_n^{*(0)}\right)^2, \tag{4}$$

where $x_n^{*(0)}$ is initial truth for entity n (such as the mean, median or mode of the claims on entity n), $|N_s|$ is the number of claims made by source s. Another interpretation of Eq.(4) is that $\hat{\sigma_s^2}$ represents the mean of the squared loss of the errors that source s makes.

However, this estimator is not precise when $|N_s|$ is very small, so it can not accurately reflect the real variance of the source. As we observed the long-tail phenomenon in Section 2, most of the sources have very few claims. Then estimator $\hat{\sigma_s^2}$ may lead to an inappropriate weight assignment for most of the sources, and further cause inaccurate truth computation. In order to solve this problem brought by the long-tail phenomenon in the dataset, we should not only consider a single value of the estimator $\hat{\sigma_s^2}$ for each source, but a range of values that can act as good estimates of σ_s^2. Therefore, we adopt the $(1-\alpha)$ confidence interval for σ_s^2, where α, also known as significant level, is usually a small number such as 0.05.

As we illustrate above, the difference between x_n^s and $x_n^{*(0)}$ follows a Gaussian distribution $N(0, \sigma_s^2)$. Since the sum of squares of standard Gaussian distribution has chi-squared distribution [17], we have:

$$\frac{\sum_{n\in N_s}\left(x_n^s - x_n^{*(0)}\right)^2}{\sigma_s^2} = \frac{|N_s|\hat{\sigma_s^2}}{\sigma_s^2} \sim \chi^2(|N_s|).$$

Thus we have:

$$P\left(\chi^2_{(1-\alpha/2,|N_s|)} < \frac{|N_s|\hat{\sigma_s^2}}{\sigma_s^2} < \chi^2_{(\alpha/2,|N_s|)}\right) = 1-\alpha,$$

which gives the $(1-\alpha)$ confidence interval of σ_s^2 as:

$$\left(\frac{\sum_{n\in N_s}\left(x_n^s - x_n^{*(0)}\right)^2}{\chi^2_{(1-\alpha/2,|N_s|)}}, \frac{\sum_{n\in N_s}\left(x_n^s - x_n^{*(0)}\right)^2}{\chi^2_{(\alpha/2,|N_s|)}}\right) \tag{5}$$

Comparing with Eq.(4), Eq.(5) is more informative. Although two sources with different numbers of claims may have the same $\hat{\sigma_s^2}$, the confidence interval of σ_s^2 for these two sources can be significantly different as shown in the following example.

Table 5: Example on calculating confidence interval

Source ID	# Claims	$\hat{\sigma_s^2}$	Confidence Interval (95%)
Source A	200	0.1	(0.0830, 0.1229)
Source B	200	3	(2.4890, 3.6871)
Source C	2	0.1	(0.0271, 3.9498)
Source D	2	3	(0.8133, 118.49)

EXAMPLE 2. *Suppose from Example 1 we obtain the statistics and sample variance for source A, B, C, and D as shown in Table 5. Both source A and C have the same $\hat{\sigma_s^2} = 0.1$, but source C has only 2 claims while source A makes 200 claims. The confidence interval of source C shows that the $\hat{\sigma_s^2}$ is rather random and the real variance may be much bigger than the sample variance for the small sources. In contrast, the confidence interval for source A is tight, and the upper bound of its confidence interval in this case is close to its $\hat{\sigma_s^2}$. Similarly, source B and D provide different numbers of claims, but they have the same $\hat{\sigma_s^2} = 3$. These two sources are not as reliable as source A and C because the sample variances are bigger, which indicates that claims made by these two sources are far from the truths. The confidence intervals for source B and D show similar patterns as source A and C. It is clear from this simple example that the confidence interval of σ_s^2 carries more information than $\hat{\sigma_s^2}$, and thus this confidence interval is helpful to estimate more accurate source weights.*

We propose to use the upper bound of the $(1-\alpha)$ confidence interval (denoted as u_s^2) as an estimator for σ_s^2 instead of using $\hat{\sigma_s^2}$ in the optimization problem 3. The intuition behind this choice is that we want to minimize the variance of $\epsilon_{combine}$ by considering the possibly worst scenario of σ_s^2 for a given source, i.e., minimize the maximum possible loss. The upper bound u_s^2 is a biased estimator on σ_s^2, but the bias is big only on sources with few claims. As the number of claims from a source increases, the bias drops.

We can substitute the unknown variance σ_s^2 in Eq.(3) by this upper bound u_s^2 and rewrite the optimization problem Eq.(3) as:

$$\begin{aligned} \min_{\{w_s\}} \quad & \sum_{s\in\mathcal{S}} w_s^2 u_s^2 \\ \text{s.t.} \quad & \sum_{s\in\mathcal{S}} w_s = 1, w_s \geqslant 0, \forall s \in \mathcal{S}. \end{aligned} \tag{6}$$

This optimization problem is convex, so the global minimum guarantees that we can find the best weight assignment under this scenario [6]. The closed form solution is:

$$w_s \propto \frac{1}{u_s^2} = \frac{\chi^2_{(\alpha/2,|N_s|)}}{\sum_{n\in N_s}\left(x_n^s - x_n^{*(0)}\right)^2}. \quad (7)$$

The weight computation (Eq.(7)) indicates that a source's weight is inversely proportional to the upper bound of the $(1-\alpha)$ confidence interval for its real variance. In Eq.(7), the chi-squared probability value will dominate the weight when a source only provides very few claims; on the other hand, if a source provides sufficient claims, the chi-squared probability value is close to $|N_s|$ and has small bias on the estimator. In this way, the proposed method automatically adjusts weights for sources with different numbers of claims. The following example illustrates the weight computation.

Table 6: Example on calculating source weight

Source ID	$\hat{\sigma_s^2}$	u_s^2	Source Weight (based on $\hat{\sigma_s^2}$)	Source Weight (based on u_s^2)
Source A	0.1	0.1229	0.4839	0.9385
Source B	3	3.6871	0.0161	0.0313
Source C	0.1	3.9498	0.4839	0.0292
Source D	3	118.49	0.0161	0.0010

EXAMPLE 3. *Based on Example 2, we compute source weights and show the results in Table 6. If we calculate the source weights based on $\hat{\sigma_s^2}$, source A and C have the same high weights because they both have low sample variances. However, if we calculate the source weights based on u_s^2, source A has a higher weight than source C. The latter weight assignment is more reasonable because source C, which provides insufficient amount of claims, may have a real variance that is much bigger than the sample variance, and therefore, should not be assigned a high weight; source A, on the other hand, may have a real variance that is close to the sample variance, so it is worth having a high weight. Similarly, source B should have a higher weight than source D. Comparing source A and B, since source A is more reliable than source B, source A has a higher weight than source B. Similarly, source C has a higher weight than source D. This example demonstrates that the upper bound of confidence interval successfully incorporates both source reliability degrees and the number of claims made by a source. Therefore it can give more accurate source weight estimation.*

3.2.3 Algorithm Flow

So far we have described how to compute the truths and source weights. Here we summarize the overall flow for the proposed CATD method in the following two steps:

Step I: Computing Source Weights. With the initial truths $\{x_n^{*(0)}\}$, we first compute the source weights based on Eq.(7).

Step II: Computing Truths. At this step, we have the weight w_s of each source, and we compute the truth for each entity by the weighted combination of the claims as shown in Eq.(1).

The pseudo code of the proposed CATD method is shown in Algorithm 1. Note that outliers should be removed before applying the proposed CATD method as the truth computation may be sensitive to outliers.

3.2.4 Practical Issues and Time Complexity

Here we discuss three techniques to make the proposed method more general and practical. At the end of this section, we analyze the time complexity of the proposed CATD method.

Algorithm 1 : **CATD**

Input: set of claims $\mathcal{C}$, significance level α.
Output: set of truths $\mathcal{X} = \{(n, x_n^*)\}$.

1: Remove outliers in claims;
2: Initialize truths $\{x_n^{*(0)}\}$;
3: Compute source weights $\{w_s\}$ according to Eq.(7) based on the initial truths $\{x_n^{*(0)}\}$;
4: **for** each $n \in \mathcal{N}$ **do**
5: Update the truth x_n^* according to Eq.(1) based on the estimation of source weights;
6: **end for**
7: **return** (n, x_n^*) pairs;

In the proposed CATD method, we use Gaussian distribution to describe the error of each source. For continuous data, the error is defined as $x_n^s - x_n^*$. For categorical data, we compute the error in the following way. We propose to represent categorical data using vectors. For example, for an answer to a multiple choice question (say four choices), choice "A" can be coded as $(1,0,0,0)$, choice "B" can be coded as $(0,1,0,0)$, and etc. Formally, if entity n has m possible values and x_n^s is the l-th value, then the claim vector $\vec{x}_n^s$ for x_n^s is defined as $\vec{x}_n^s = (0,\ldots,\overset{l}{1},0,\ldots,0)^T$. In this way, a categorical type claim is represented by a vector. In order to compute the variance in Eq.(4), we can use the square of L^2-norm to indicate the difference between the claim vectors and the initial truth vectors.

Another important issue is the scale of entities. As illustrated in the weight computation, we model the errors as Gaussian distributed. If different entities have significantly different scales, the errors on the entity that has larger scale may be significantly bigger than errors on other entities. To solve this issue, we can normalize the claims of the same entity so that the scale on all the entities falls into the same range.

In Algorithm 1, source weights are estimated only once and then the truths are computed as weighted combination of the claims. To achieve a more accurate result, we can adopt an iterative procedure on both source weights and truth computation. In each iteration, the proposed CATD method improves the variance estimation using the latest truths, so that the weight assignment is more appropriate. Then the truths are updated based on the current weight assignment. We stop the procedure until termination criterion is met, which can be set as the maximum number of iterations or a threshold for the similarity between truths from current computation and the truths from the previous computation.

The time complexity of the CATD method is linear with respect to the total number of claims, i.e. $O(|\mathcal{C}|)$, where $|\mathcal{C}|$ is the input size of the proposed method. If the aforementioned iterative procedure is adopted, the time complexity of the CATD method is then changed to $O(|\mathcal{C}| * m)$, where m is the number of iterations. The time complexity is experimentally validated in Section 4.5.

4. EXPERIMENTS

In this section, we test the proposed CATD method on four real world applications and various simulated datasets. The experimental results show that the proposed method outperforms state-of-the-art truth discovery methods when confronting the challenge that data present long-tail phenomenon. We first discuss the experiment setup in Section 4.1, and then validate our assumption that source

errors are Gaussian distributed in Section 4.2. We present experimental results on the aforementioned City Population, Biography and Indoor Floorplan datasets in Section 4.3, and on Game dataset in Section 4.4. Finally in Section 4.5, the proposed method is tested on different scenarios of simulated datasets.

4.1 Experiment Setup

In this part, we introduce the performance measures for different data types and the baseline methods.

Performance Measures

In the experiments, we have two types of data: continuous and categorical. To evaluate the performance of various truth discovery methods, we adopt the following measures for these two data types:

- *MAE*: For continuous data, we report the mean of absolute error (*MAE*) which measures the mean of absolute distance from the approach's output to the ground truths.
- *RMSE*: For continuous data, we also report the root of mean squared error (*RMSE*). This measurement penalizes more on the large distance and less on the small distance comparing with MAE.
- *Error Rate*: For categorical data, we measure the *Error Rate* by computing the percentage of mismatched values between each approach's output and ground truths. For continuous data, we also report error rate, where a "mismatch" is defined as the case when the distance from the ground truth is greater than a threshold (e.g. 0.1% of the ground truth values).

As a lower measurement value means that the method's estimation is closer to the ground truths, for all measures, the **lower** the value, the **better** the performance.

Baseline Methods

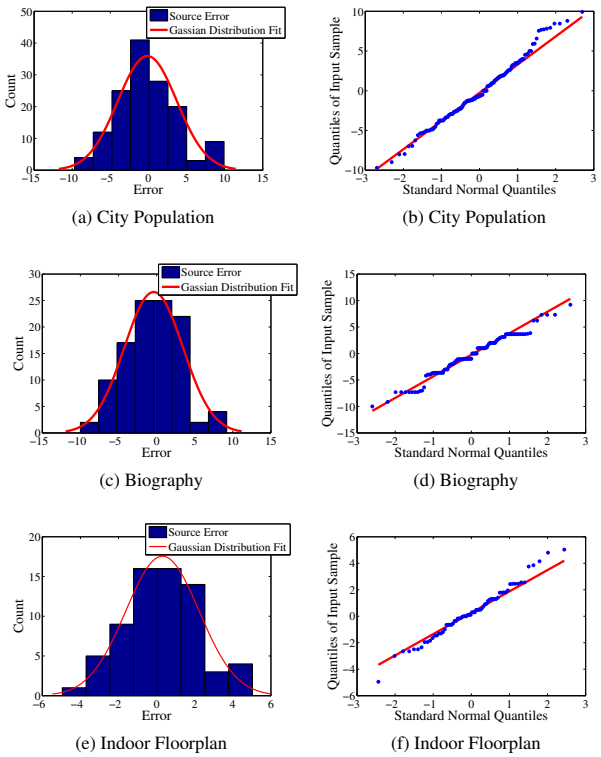
All baseline methods that we use in the experiments are conducted on the same input set in an unsupervised manner. The ground truths are used only in evaluation. They are compared on both continuous and categorical data unless otherwise specified. The baseline methods include some state-of-the-art truth discovery methods: *GTM* [36], *TruthFinder* [34], *AccuSim* [8], *Investment* [23], *3-Estimates* [16], and *CRH* [18]. More detailed summary of these methods can be found in Section 5. We also compare with naive conflict resolution methods: *Mean* (the truth for each entity is the mean of the claims), *Median* (the truth for each entity is the median of the claims), and *Voting* (the truth for each entity is the claim that stated by the most sources). Note that *GTM*, *Mean*, and *Median* only apply to continuous data type and is not used in the experiments on Game dataset (categorical data type).

Note that some algorithms have extended versions which take into account source dependency, but they are not compared in the experiment because we do not consider source dependency in this paper but leave it for future work. Parameters for the above methods are set according to the suggestions by their authors. For the proposed method, iterative procedure is applied and the significance level α is set as 0.05.

4.2 Assumption Validation

Since we have Gaussian assumption on source error distribution, we first conduct normality tests to validate this assumption. We use the aforementioned City Population, Biography and Indoor Floorplan datasets for this purpose.

Figure 3 shows the error distributions (left column) of sources from City Population, Biography and Indoor Floorplan datasets respectively. Gaussian distributions are fitted and the mean is approximate 0. We use Q-Q plot, a well-known graphical technique



(a) City Population (b) City Population

(c) Biography (d) Biography

(e) Indoor Floorplan (f) Indoor Floorplan

Figure 3: Source error distributions.

for normality testing, to further validate the error distributions. In Q-Q plots (right column), data points are plotted against a theoretical Gaussian distribution (the line in the plot) and an approximate straight line indicates strong normality. Figure 3 proves that the source errors are indeed Gaussian distributed.

4.3 Experiments on Long-Tail Datasets

In this section, we show the experimental results on three real world datasets: City Population dataset, Biographic dataset and Indoor Floorplan dataset, which contain numerical data. Some discussions about these datasets can found be in Section 2. Comparing with the baseline methods, the proposed CATD method shows the power of finding more accurate truths on datasets with long-tail phenomenon.

City Population Dataset [23, 36]. City Population dataset contains Wikipedia edit histories of some cities' population in a given year. A subset of the entities (308 out of 43,071) are randomly sampled and labeled with ground truths. The following preprocessing steps are applied on this dataset: First, since the data is Wikipedia edit history, a source may have multiple claims on the same entity. Considering sources may update their claims to correct previous errors, we only keep the latest claims of each source on one entity. Second, we observe some unreasonable claims such as 0 and 6.5979×10^{18} for a city's population. Therefore, we adopt the same preprocessing from [36].

Biography Dataset [23, 36]. This dataset contains Wikipedia edit histories about people's biographic information. Similar to City Population dataset, the ground truths are available for some entities (2,685 out of 9,924) and the same data preprocessing is conducted. Note that City Population and Biography datasets (both data and ground truths) are provided by the authors of [23, 36] and more details about the data can be found in the papers.

Indoor Floorplan Dataset. Indoor Floorplan dataset contains the distance estimates from users' smart phones for indoor hallways. We manually label the ground truths of each hallways distance (129 out of 129) by measuring tapes.

In Table 7, 8, and 9 we summarize the performance of all methods in terms of MAE, RMSE and Error Rate. We can see that the proposed CATD method achieves the best performance on all datasets comparing with the baseline methods.

Table 7: Comparison on City Population dataset

Method	MAE	RMSE	Error Rate
CATD	**1203.28**	**8075.56**	**0.1423**
Mean	10368.54	126199.76	0.6058
Median	10241.81	126198.86	0.3577
Voting	10327.20	126217.98	0.5255
GTM	1498.59	8339.99	0.1606
TruthFinder	1633.60	8824.09	0.1715
AccuSim	1626.52	8718.10	0.1642
Investment	1617.40	8797.43	0.1679
3-Estimates	1640.83	8822.50	0.1569
CRH	1425.46	8569.44	0.1679

Table 8: Comparison on Biography dataset

Method	MAE	RMSE	Error Rate
CATD	**211.54**	**4727.36**	**0.0361**
Mean	253.41	4860.28	0.0528
Median	244.04	4854.90	0.0377
Voting	237.35	4847.80	0.0366
GTM	228.19	4831.53	0.0366
TruthFinder	278.08	4905.71	0.0371
AccuSim	267.62	4844.46	0.0366
Investment	369.04	5380.71	0.0439
3-Estimates	237.35	4847.80	0.0366
CRH	244.17	4832.20	0.0377

Table 9: Comparison on Indoor Floorplan dataset

Method	MAE	RMSE	Error Rate
CATD	**0.9960**	**1.3845**	**0.1240**
Mean	1.7851	2.2846	0.3488
Median	1.3797	1.7860	0.2326
Voting	1.6029	2.1153	0.3023
GTM	1.2845	1.6823	0.2403
TruthFinder	1.4754	2.0467	0.2713
AccuSim	1.3964	1.9191	0.2481
Investment	1.7243	2.5803	0.3256
3-Estimates	1.7417	2.6075	0.3101
CRH	1.1929	1.5955	0.1783

On City Population dataset, there are many entities with very few claims, which makes Mean, Median and Voting's performance not satisfactory. The truth discovery methods can achieve better performance than the naive baselines. TruthFinder and AccuSim achieve nice results by additionally considering the influence between claims. CRH and GTM, which take into account the characteristic of continuous data for truth discovery tasks, also obtains good results. CATD method makes further improvement in terms of MAE, RMSE and Error Rate by taking into account the long-tail phenomenon in the dataset. The improvement is over 18% on MAE, 3% on RMSE, and 10% on Error Rate comparing with the best baselines.

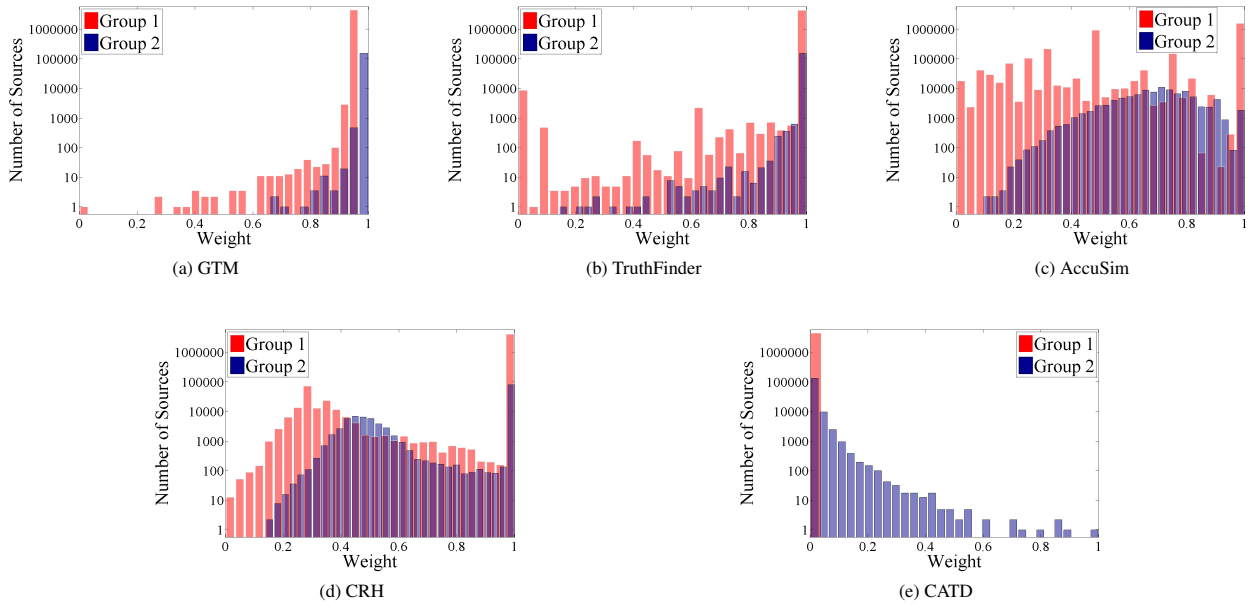
On Biography dataset, since there are more claims for each entity in general, Mean, Median and Voting perform well even without considering source reliability. Truth discovery methods that ignores the impact of long-tail phenomenon in the data cannot learn the source reliability correctly, and thus obtains similar or even worse results comparing to the naive baselines. Among all the baseline methods, GTM provides competitive results, but the proposed CATD method further improves the results in terms of MAE, RMSE and Error Rate by 7.9%, 2.2% and 1.4% respectively by considering the long-tail phenomenon in the dataset.

On Indoor Floorplan dataset, the proposed CATD method consistently provides the best results in terms of MAE, RMSE and Error Rate comparing with other baseline methods. The improvement over the best baseline method is 19.8% on MAE, 15.3% on RMSE, and 43.8% on Error Rate. By outperforming the baseline methods on all real world datasets, the proposed CATD method demonstrates its power on modeling source reliability accurately even when the sources make insufficient claims.

Since all the truth discovery methods and the proposed CATD method use weighted voting or averaging to calculate truths, the estimated source reliability is the key to obtain accurate truths. Therefore, we further examine the source reliability degrees on Biography dataset given by each method in the following.

As different methods adopt various weight computation, we normalize the source weights into the range $[0, 1]$ by dividing the maximum weight to make a fair comparison. In order to illustrate the problem brought by the long-tail phenomenon, sources are divided into two groups: Group 1 contains sources with less than five claims and Group 2 contains sources with five or more claims. This threshold is set so that the ratio of group sizes is not too extreme.

Intuitively, Group 1 sources should have small weights, because each of them provides only few claims. Group 2 sources may have large weights or small weights depending on the sources' reliability. Figure 4 shows the weight distributions of these two groups of sources for GTM, TruthFinder, AccuSim and CRH baseline methods and the proposed CATD method. We choose these three baselines here because the other truth discovery baselines are designed for categorical data only, so the weights learned by those methods on numerical claims are not representative. As we can see in the figures, GTM distinguishes Group 1 and Group 2 sources to some extent, but due to the setting of the prior on source reliability, the difference is small and it overestimates the reliability degrees for Group 1 sources. The problem for TruthFinder and AccuSim is that the weight distribution of Group 1 sources is polarized. The number of Group 1 sources which have weights as high as 1 stands out. Each Group 1 source only makes a few claims, and if the claims are correct, then its accuracy is high, so TruthFinder and AccuSim assign a large weight to this source. If the few claims are wrong, the corresponding source's accuracy is low, so it is assigned a small weight. Although TruthFinder and AccuSim have reasonable source reliability estimation on big sources, the inaccurate estimation on large amount of small sources discounts their performance. The same observation can be found on CRH, which also ignores the difference between big and small sources, and assign source weights purely based on accuracy without considering the sample size. Only the proposed CATD method is aware that when the claims made by a small source happen to be accurate, it does not confirm that this small source is reliable; and for big sources, the bias on source reliability estimation is low. From Figure 4e, we can see that Group 1 sources have relatively low weights. For Group 2 sources, some of them have low weights whereas others have big weights. Thanks to the accurate source reliability estimation, the proposed CATD method provides more accurate truths.



(a) GTM (b) TruthFinder (c) AccuSim

(d) CRH (e) CATD

Figure 4: Comparison of source weights on Biography dataset.

Table 10: Comparison on Game dataset

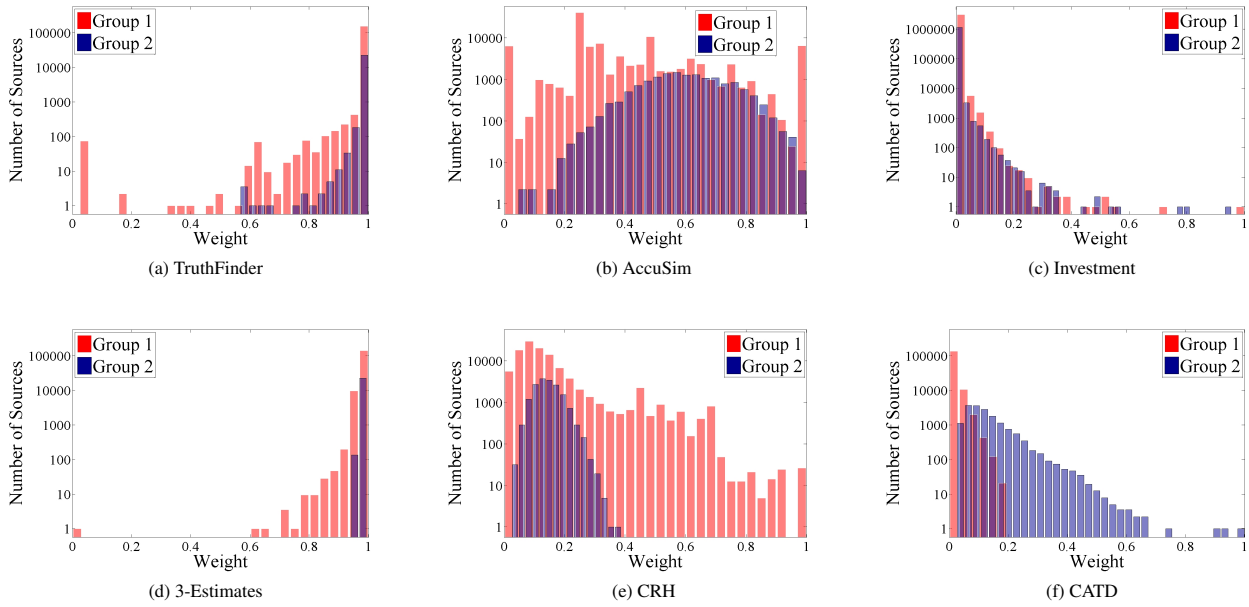
Method	Error Rate										
	Level 1 (303)	Level 2 (295)	Level 3 (290)	Level 4 (276)	Level 5 (253)	Level 6 (218)	Level 7 (187)	Level 8 (138)	Level 9 (99)	Level 10 (44)	Overall (2103)
CATD	**0.0132**	**0.0271**	**0.0276**	**0.0290**	**0.0435**	**0.0596**	**0.0481**	**0.1304**	**0.1414**	**0.2045**	**0.0485**
Voting	0.0297	0.0305	0.0414	0.0507	0.0672	0.1101	0.1016	0.3043	0.3737	0.5227	0.0980
TruthFinder	0.0693	0.0915	0.1241	0.0942	0.1581	0.2294	0.2674	0.3913	0.5455	0.5455	0.1816
AccuSim	0.0264	0.0305	0.0345	0.0507	0.0632	0.0963	0.0909	0.2826	0.3636	0.5000	0.0913
Investment	0.0330	0.0407	0.0586	0.0761	0.0870	0.1239	0.1283	0.3406	0.3838	0.5455	0.1151
3-Estimates	0.0264	0.0305	0.0310	0.0507	0.0672	0.1055	0.0963	0.2971	0.3737	0.5000	0.0942
CRH	0.0264	0.0271	0.0345	0.0435	0.0593	0.0872	0.0856	0.2609	0.3535	0.4545	0.0866

4.4 Experiments on Game Dataset

In this section, we test the generalizability of the proposed CATD method on discrete data. Game dataset [2] collects multi-source answers based on a TV game show, details of which can be found in Section 2. Different from the other three datasets used in this paper, Game dataset contains categorical type of claims. Therefore, we encode the claims into vectors and then apply CATD method. We use error rate as the evaluation metric on this dataset.

The ground truth information is provided by the TV game show. In addition, the show gives each question a difficulty level. We use the ground truth information for 2103 out of 2169 questions. The remaining questions' difficulty levels are missing and are excluded in the evaluation. Table 10 shows the number of ground truths for each level of questions (number in parentheses) and error rate of the proposed CATD method and baseline methods on all question levels. We can see that CATD method is more accurate than the baseline methods on every level. Overall, we reduce the error rate by almost half comparing with the state-of-the-art truth discovery methods. The first seven levels have low error rates on all methods because those questions are relatively easy; for the last three levels, as the questions get harder, the error rates for all baseline methods increase dramatically. The error rates for the proposed CATD method also increase slightly on the last three levels, but show a large advantage over all baseline methods. On the hardest question level, CATD method still has error rate as low as 0.2045, while the baselines have error rates greater than 0.4. TruthFinder and Investment perform worse than Voting because both methods model the probability of each claim being correct given the source reliability degrees without considering complement vote. However, under this application scenario, complement vote should be considered because if a player votes for choice A, it naturally means he/she votes against other choices.

Similar to the source reliability analysis that we conduct on Biography dataset, we explore the weight distribution of Game dataset on two groups of sources: Group 1 contains sources with fewer than 10 claims and Group 2 contains sources with 10 or more claims. The threshold is set so that the group sizes are comparable. Figure 5 shows the weight distributions of these two groups for the truth discovery baseline methods and the proposed CATD method. We can see that TruthFinder and AccuSim have many Group 1 sources with weights either very big (close to 1) or very small (close to 0), which presents a similar polarization distribution as they present on Biography dataset. Because both methods apply Bayesian analysis, the effect of long-tail phenomenon causes similar problem when estimating source reliability on small sources. Investment method presents the same weight distribution on both groups, which indicates that they do not realize that the number of claims made by each source can be an influential factor for source reliability estimation. 3-Estimates is too optimistic and overestimates the source reliability degrees on both groups. Although TruthFinder and 3-



(a) TruthFinder (b) AccuSim (c) Investment

(d) 3-Estimates (e) CRH (f) CATD

Figure 5: Comparison of source weights on Game dataset.

Estimates present similar pattern of weight distribution, 3-Estimate does a slightly better job on distinguishing Group 1 and Group 2 sources, which leads to a better performance than TruthFinder. CRH overestimates the source reliability on small sources, which is caused by the long-tail phenomenon in the data. The proposed CATD method is able to take the number of claims made by a source into consideration, and obtains appropriate weight assignment. The advantage of the proposed CATD method in source weight assignment is demonstrated in Figure 5f. We can see that Group 1 sources have relatively low weights while Group 2 sources have weights distributed over the entire interval. It is desired because only when sources have sufficient claims it is meaningful to estimate their reliability; for small sources, we hope they do not affect the truth estimation heavily. That is the reason why the proposed CATD method performs better than baseline methods.

4.5 Multi-source Long-tail Simulations

In this section, we first simulate different scenarios involving various distributions of source reliability and test the effectiveness and necessity of the confidence interval estimator in the proposed method. Last but not least, we use the simulated data to test the efficiency of the proposed CATD method.

We generate datasets containing 200 entities. Each source chooses a subset of entities randomly and makes claims. Here we use power law function to create the long-tail phenomenon and the parameter is set as $\#sources = \#claims^{-1.5} \cdot e^7$. Totally 2712 sources are generated and for each source, we generate claims from a $N(\mu, \sigma_s^2)$ distribution and set the ground truths for all entities as μ. In the following experiments, we simulate four different scenarios which can be found commonly in real world applications by varying source reliability distributions, i.e., distributions of σ_s^2. To reduce the randomness from a single dataset, we generate the datasets and run the experiments for 100 times and report the average of the results. The measurements for the results include MAE, RMSE and Error Rate.

In order to test the effectiveness and necessity of using upper bound u_s^2 in Eq.(6), *CATD-0* is compared, where CATD-0 uses the same framework as CATD but with estimated variance $\hat{\sigma_s^2}$ (Eq.(4)) instead of the upper bound. The comparison between the proposed CATD method and this baseline can demonstrate the importance of considering source size in source reliability estimation. *CATD-0* only uses sample variance to estimate source reliability, and thus it cannot accurately estimate small sources' reliability degree. In contrast, the proposed CATD method takes sample size into consideration when estimating sources' weights via the usage of variance upper bound. Median, GTM, and CRH, the top three baselines on our simulated data, are used as a performance reference. Due to space limit, other baselines are omitted here. From the experiments, GTM and CRH show their strength when handling numerical values, but the performance of CATD is still consistently the best, which illustrates the importance of considering long-tail phenomenon in truth discovery tasks.

Scenario 1: $\sigma_s^2 \sim$ Uniform$(0, 3)$. The uniform distribution of source reliability implies that we have equal amount of sources with various reliability degrees. We can see from Table 11 that CATD improves CATD-0 by taking the long-tail phenomenon into consideration.

Table 11: Comparison on simulated datasets: Scenario 1

Method	MAE	RMSE	Error Rate
CATD	0.0287	0.0365	0.0117
CATD-0	0.0471	0.0657	0.1082
Median	0.0525	0.0709	0.1381
GTM	0.0434	0.0546	0.0684
CRH	0.0454	0.0619	0.0962

Scenario 2: $\sigma_s^2 \sim$ Folded Normal$(\mu = 1, \sigma^2 = 1)$. This distribution generates more reliable sources than unreliable ones. Comparing with Scenario 1, the unreliable sources here have a larger variance. Table 12 shows the results that CATD method still has a big advantage over other methods, whereas CATD-0 loses its advantage over Median.

Scenario 3: $\sigma_s^2 \sim$ Gamma$(k = 1, \theta = 1.5)$. Under this setting, the Gamma distribution is equivalent to an exponential distribution,

Table 12: Comparison on simulated datasets: Scenario 2

Method	MAE	RMSE	Error Rate
CATD	0.0217	0.0274	0.0012
CATD-0	0.0369	0.0518	0.0647
Median	0.0373	0.0507	0.0557
GTM	0.0312	0.0391	0.0125
CRH	0.0319	0.0438	0.0330

which implies that most of the sources are reliable with small variances, whereas a few sources are very unreliable with large variances. Table 13 shows the results for this scenario. Note that the performance of Median is now better than CATD-0. It is because that when there are some very unreliable sources, CATD-0 may underestimate the variances of those sources, and thus lead to inappropriate weight estimation. CATD method, on the other hand, is not affected and outperforms other methods.

Table 13: Comparison on simulated datasets: Scenario 3

Method	MAE	RMSE	Error Rate
CATD	0.0228	0.0289	0.0014
CATD-0	0.0463	0.0674	0.1084
Median	0.0309	0.0427	0.0305
GTM	0.0354	0.0444	0.0268
CRH	0.0253	0.0354	0.0131

Scenario 4: $\sigma_s^2 \sim \text{Beta}(\alpha = 0.5, \beta = 0.5)$. The Beta distribution that we choose has a "U" shape, which implies that most of the sources are either reliable with small variances or very unreliable with large variances. Table 14 shows the results for this scenario. Note that the performance of Median is better than CATD-0, and the reason is the same as we describe in Scenario 3. CATD method is still the best and outperforms the other methods.

Table 14: Comparison on simulated datasets: Scenario 4

Method	MAE	RMSE	Error Rate
CATD	0.0507	0.0668	0.1168
CATD-0	0.1059	0.1695	0.3349
Median	0.0562	0.0953	0.1823
GTM	0.1440	0.1809	0.5753
CRH	0.0656	0.0987	0.1658

The experiments on simulated datasets demonstrate that if long-tail phenomenon in the data is ignored and small sources are treated the same as big sources, the estimation of source reliability based on sample variance cannot help or even harm the performance. Estimator based on the confidence interval of source reliability in CATD helps the method perform robustly under different source reliability distributions on long-tail data, because it takes both source reliability and source size into consideration.

Efficiency is an important aspect of truth discovery tasks. Here we generate different numbers of claims for the simulated data and test the computational complexity of the CATD method. Figure 6 shows strong linearity between the running time and the number of claims. To further prove the linearity, we compute Pearsons correlation coefficient, a commonly used metric to test linear relationship between variables. The closer it is to 1 (or -1), the stronger positive (or negative) linear relationship the variables have. In our experiment, the Pearsons correlation coefficient for running time and the number of claims is 0.9991, indicating that they are highly linearly correlated.

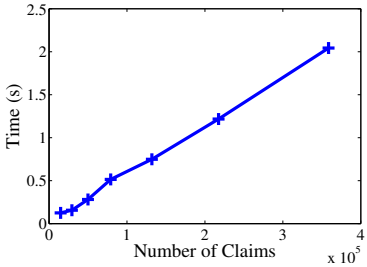


Figure 6: Running time of CATD w.r.t. number of claims.

5. RELATED WORK

Improving data quality has been studied in the database community for years [12, 13]. As an important aspect of this area, research on resolving conflicts from multiple sources [4, 5, 10, 22] arise various ways to handle conflicts in data integration. A common method is to conduct voting or averaging–for categorical data, the information with the highest number of occurrences is regarded as truth; for continuous claims, the mean is taken as the true value. This approach regards sources as equally reliable, and thus may fail when there exist sources containing low quality information.

To address the problem voting or averaging may face, truth discovery has received lots of attention recently and many methods are developed to handle various challenges. The truth discovery problem is first formally formulated by Yin et al. [34], in which a Bayesian based heuristic algorithm is proposed. It computes the probability of each claim being correct given the estimated source weights and the influences between claims. Then Pasternack et al. propose methods to incorporate prior knowledge, such as constraints on truth and background information, into truth discovery tasks [23, 24]. The methods adopt the idea that sources "invest" their reliability on the claims they provide. The concept of difficulty of getting the truth when computing source weights is modeled in [16], which also adopts the idea of complement vote. Algorithms proposed in [8] targeted to handle source dependency problem in truth discovery tasks. The basic idea is that if two sources make the same incorrect claims, they are likely to be correlated. Bayesian analysis and the idea of complement vote are adopted. Note that the method compared in Section 4 does not consider source dependency, but considers the influences between claims. A semi-supervised graph learning is proposed in [35]. It models the propagation of information trustworthiness from the known ground truths. Zhao et al. adopt probabilistic graphical models in truth discovery tasks [36, 37]. The existence of multiple truths for single entity is considered in [37] where source reliability is modeled as two-sided: sensitivity and specificity. Later, a model specially designed for numerical data is proposed in [36]. Recently, Dong et al. model source selection in the truth discovery tasks based on the idea of "gain" and "cost" [11, 27]. Li et al. aim to minimize the weighted deviation of claims and truths, so an optimization framework is adopted and applied on heterogeneous data, in which different data types can be modeled jointly [18].

Another related field is learning from crowd of wisdom, also known as crowdsourcing, in which researchers investigate how to infer true labels from the labeling efforts of a crowd [3, 15, 29, 30, 32, 33, 38]. These approaches focus on learning true labels or answers to certain questions, where the input and output space are usually limited to specific sets of labels or questions. Truth discovery tasks, on the other hand, can deal with more than categorical data type comparing with crowdsourcing tasks.

Long-tail phenomenon has attracted attentions in many fields [7, 14, 21] for theoretical and practical reasons. However, it is not

identified by any existing truth discovery work. To the best of our knowledge, we are the first to analyze the effect of this ubiquitous phenomenon and propose an effective solution.

6. CONCLUSIONS

Truth discovery is important for effective data integration because it can automatically identify reliable sources and trustworthy information from multiple sources. We observe long-tail phenomenon in many real world applications, i.e., most sources only provide very few claims and only a small amount of sources makes plenty of claims. This phenomenon causes the problem that the existing work cannot appropriately estimate source reliability from insufficient information. In this paper, we propose a confidence-aware truth discovery (CATD) method to resolve the conflict on data with long-tail phenomenon by adopting effective estimators based on the confidence interval of source reliability. These estimators can successfully discount the effect of small sources and accurately reflect the real source reliability. Experiments are conducted on four real world applications as well as simulated datasets with various source reliability distributions. The results demonstrate an advantage of the proposed CATD method over existing truth discovery approaches in finding truths on long-tail data.

7. ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their valuable comments and suggestions, which help us tremendously in improving the quality of the paper. We also thank Si Chen for his help in data collection. The work was supported in part by National Key Basic Research Program of China (973 Program) under Grant No. 2014CB340304, the National Science Foundation under Grant NSF IIS-1319973, the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), and the U.S. Army Research Office under Cooperative Agreement No. W911NF-13-1-0193.

8. REFERENCES

[1] M. Alzantot and M. Youssef. Crowdinside: Automatic construction of indoor floorplans. In *Proc. of SIGSPATIAL*, pages 99–108, 2012.

[2] B. I. Aydin, Y. S. Yilmaz, Y. Li, Q. Li, J. Gao, and M. Demirbas. Crowdsourcing for multiple-choice question answering. In *IAAI*, pages 2946–2953, 2014.

[3] Y. Bachrach, T. Minka, J. Guiver, and T. Graepel. How to grade a test without knowing the answers – a bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *Proc. of ICML*, pages 255–262, 2012.

[4] J. Bleiholder and F. Naumann. Conflict handling strategies in an integrated information system. In *Proc. of WWW*, 2006.

[5] J. Bleiholder and F. Naumann. Data fusion. *ACM Computing Surveys*, 41(1):1:1–1:41, 2009.

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

[7] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

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

[9] X. L. Dong, L. Berti-Equille, and D. Srivastava. Data fusion: resolving conflicts from multiple sources. In *Handbook of Data Quality*, pages 293–318. Springer, 2013.

[10] X. L. Dong and F. Naumann. Data fusion: Resolving data conflicts for integration. *PVLDB*, 2(2):1654–1655, 2009.

[11] X. L. Dong, B. Saha, and D. Srivastava. Less is more: Selecting sources wisely for integration. *PVLDB*, 6(2):37–48, 2012.

[12] W. Fan. Data quality: Theory and practice. *Web-Age Information Management*, page 1–16, 2012.

[13] W. Fan, F. Geerts, S. Ma, N. Tang, and W. Yu. Data quality problems beyond consistency and deduplication. *In Search of Elegance in the Theory and Practice of Computation*, pages 237–249, 2013.

[14] R. Feldman and M. Taqqu. *A practical guide to heavy tails: statistical techniques and applications*. Springer, 1998.

[15] J. Feng, G. Li, H. Wang, and J. Feng. Incremental quality inference in crowdsourcing. In *Database Systems for Advanced Applications*, pages 453–467. Springer, 2014.

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

[17] R. V. Hogg, J. McKean, and A. T. Craig. *Introduction to mathematical statistics*. Pearson Education, 2005.

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

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

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

[21] E. Mustafaraj, S. Finn, C. Whitlock, and P. T. Metaxas. Vocal minority versus silent majority: Discovering the opionions of the long tail. In *Proc. of IEEE SocialCom*, pages 103–110, 2011.

[22] F. Naumann, A. Bilke, J. Bleiholder, and M. Weis. Data fusion in three steps: Resolving schema, tuple, and value inconsistencies. *IEEE Data Engineering Bulletin*, 29(2):21–31, 2006.

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

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

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

[26] G.-J. Qi, C. C. Aggarwal, J. Han, and T. Huang. Mining collective intelligence in diverse groups. In *Proc. of WWW*, pages 1041–1052, 2013.

[27] T. Rekatsinas, X. L. Dong, and D. Srivastava. Characterizing and selecting fresh data sources. In *Proc. of SIGMOD*, pages 919–930, 2014.

[28] G. Shen, Z. Chen, P. Zhang, T. Moscibroda, and Y. Zhang. Walkie-markie: indoor pathway mapping made easy. In *Proc. of NSDI*, pages 85–98, 2013.

[29] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proc. of KDD*, pages 614–622, 2008.

[30] Y. Tian and J. Zhu. Learning from crowds in the presence of schools of thought. In *Proc. of KDD*, pages 226–234, 2012.

[31] V. Vydiswaran, C. Zhai, and D. Roth. Content-driven trust propagation framework. In *Proc. of KDD*, pages 974–982, 2011.

[32] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *NIPS*, volume 10, pages 2424–2432, 2010.

[33] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. R. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *NIPS*, volume 22, pages 2035–2043, 2009.

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

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

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

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

[38] D. Zhou, J. C. Platt, S. Basu, and Y. Mao. Learning from the wisdom of crowds by minimax entropy. In *NIPS*, pages 2204–2212, 2012.