Towards Confidence in the Truth: A Bootstrapping based Truth Discovery Approach

Houping Xiao[1], Jing Gao[1], Qi Li[1], Fenglong Ma[1], Lu Su[1]
Yunlong Feng[2], and Aidong Zhang[1]
[1]SUNY Buffalo, Buffalo, NY USA [2]KU Leuven, Leuven, Belgium
{houpingx,jing,qli22,fenglong,lusu}@buffalo.edu
yunlong.feng@esat.kuleuven.be, azhang@buffalo.edu

ABSTRACT

The demand for automatic extraction of true information (i.e., truths) from conflicting multi-source data has soared recently. A variety of *truth discovery* methods have witnessed great successes via jointly estimating source reliability and truths. All existing truth discovery methods focus on providing a point estimator for each object's truth, but in many real-world applications, confidence interval estimation of truths is more desirable, since confidence interval contains richer information. To address this challenge, in this paper, we propose a novel truth discovery method (*ETCIBoot*) to construct confidence interval estimates as well as identify truths, where the bootstrapping techniques are nicely integrated into the truth discovery procedure. Due to the properties of bootstrapping, the estimators obtained by *ETCIBoot* are more accurate and robust compared with the state-of-the-art truth discovery approaches. Theoretically, we prove the asymptotical consistency of the confidence interval obtained by *ETCIBoot*. Experimentally, we demonstrate that *ETCIBoot* is not only effective in constructing confidence intervals but also able to obtain better truth estimates.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*

Keywords

Truth Discovery; Confidence Interval; Bootstrapping

1. INTRODUCTION

Today, we are living in a data-rich world, and the information on an object (e.g., population/weather/air quality of a particular city) is usually provided by multiple sources. Inevitably, there exist conflicts among the multi-source data due to a variety of reasons, such as background noise, hardware quality or malicious intent to manipulate data, etc. An important question is how to identify the true information (i.e., truths) among the multiple conflicting pieces of information. Because of the volume issue, we cannot expect people to detect truth for each object manually. Thus, the demand for automatic extraction of truths from conflicting multi-source data has soared recently.

A commonly used multi-source aggregation strategy is averaging or voting. The main drawback of these approaches is that they treat the reliability of each source as the same. In real-world applications, however, different sources may have different degrees of reliability and more importantly, their reliability degrees are usually unknown *a priori*. To address this problem, a variety of truth discovery methods [2–5, 7, 11, 12, 15–18, 20–25] have been proposed. Although these methods vary in many aspects, they share a common underlying principle: If a piece of information is provided by a reliable source, it is more likely to be trustworthy, and the source that more often provides trustworthy information is more reliable. Following this principle, existing methods are designed to jointly estimate source reliability and truths by assigning larger weights to the reliable sources which in return play more important roles in the data aggregation.

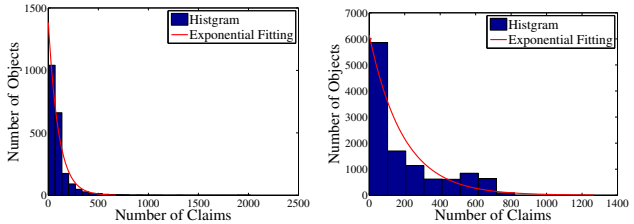
All existing truth discovery methods [2–5, 7, 11, 12, 15–19, 21, 22] focus on providing a point estimator for each object's truth, i.e., the estimate is a single value. However, important confidence information is missing in this single-value estimate. For example, two objects A and B receive the same truth estimate, e.g., 25. Even though the estimates are the same, the confidence in these estimates could differ significantly–A may receive 1000 claims around 25 while B only receives one claim of 25, and clearly the confidence in A's truth estimate is much higher. Therefore, instead of a point estimation, an estimated confidence interval of the truth is more desirable. An α-level confidence interval [8] is an interval (a, b) such that $\mathbb{P}(\theta \in (a, b)) = \alpha$ for a given $\alpha \in (0, 1)$, where θ denotes the truth in our scenario. The width of the interval reflects the confidence in the estimate–A smaller interval indicates the higher confidence in the estimate and a larger interval means that the estimate has more possible choices within the interval. In the example we just mentioned, suppose the 95% conference interval of A and B's estimates are (24.9,25.1) and (0,50) respectively. Although both truth estimates are 25, we are more certain that the truth of A is close to 25. With such confidence information, the decision makers can use the truth estimates more wisely. However, such important confidence information cannot be

KDD '16, August 13-17, 2016, San Francisco, CA, USA

DOI: http://dx.doi.org/10.1145/2939672.2939831

obtained by the traditional point estimation strategy adopted by existing truth discovery methods.

The estimation of confidence intervals for objects' truths can benefit any truth discovery scenario by providing additional information (i.e., confidence) in the output, but its advantage is more obvious on long-tail data. A multi-source data is said to be long-tail in the sense that most objects receive a few claims from a small number of sources and only a few objects receive many claims from a large number of sources. As discussed in the aforementioned example, the difference in the confidence of the truth estimates is usually caused by the difference in the number of claims received by the objects. When an object receives more claims, a smaller confidence interval is obtained, and thus the estimate of this truth is more certain. It is essential to provide confidence intervals rather than points for the truth estimates on such long-tail data, which are ubiquitous. The Flight Status and Game applications used in our experiments are examples of such long-tail phenomena (The details are deferred in Subsection 4.3). In Figure 1, we present the histograms in terms of the number of claims and fit them into an exponential distribution, a typical long-tail distribution, respectively.



(a) Flight Status Dataset (b) Game Dataset

Figure 1: *The long-tail phenomenon*

To address the problem, in this paper, we propose a novel method, **E**stimating **T**ruth and **C**onfidence **I**nterval via **Boot**strapping (*ETCIBoot*) to construct confidence interval estimates for truth discovery tasks. We adopt the iterative two-step procedure used in traditional truth discovery methods: 1) Update truth estimates based on the current estimates of source weights (source reliability degrees), and 2) update source weights based on the current estimates of truths. At the truth computation step, instead of giving a point estimation, we now adopt the following procedure to obtain confidence interval estimates. *ETCIBoot* obtains multiple estimates of an object's truth, using bootstrapping techniques. Each estimate is obtained by calculating the weighted averaging or voting on a new set of sources which are bootstrapped from available sources. A statistic T that involves the truths is constructed. Its distribution F is usually unknown *a priori*. Based on these multiple estimates obtained via bootstrapping, we derive an estimator $\widehat{T}$ of T and further approximate F by $\widehat{F}$ (i.e., the distribution of $\widehat{T}$). The confidence intervals of the truths are naturally implied in the distribution of $\widehat{T}$ (i.e., $\widehat{F}$). Theoretically, we prove that $\widehat{T}$ is asymptotically consistent to T in distribution, and the end points of the confidence intervals converge to the true ones at $O_p(n^{-\frac{3}{2}})$, where n is the number of claims.

Besides providing confidence intervals, *ETCIBoot* is also able to provide more accurate and robust truth estimates if we use the average of the multiple estimates as the point estimator. Existing truth discovery methods typically compute weighted mean in the truth computation step, and thus the truth estimates can be quite sensitive to some outlying claims. In contrast, *ETCIBoot* adopts bootstrapping procedure which improves the robustness of the estimation. The truth estimates are obtained by computing the mean of bootstrap samples. These samples capture the distribution of claims in which the outlying claims' effect can be greatly reduced.

We conduct experiments on both simulated and real-world datasets. Experimental results show that the proposed *ETCIBoot* can effectively construct confidence intervals for each objects and achieve better truth estimates compared with the state-of-the-art truth discovery methods.

To sum up, the paper makes the following contributions:

- To the best of our knowledge, we are the first to illustrate the importance of confidence interval estimation in *truth discovery*, and propose an effective method (*ETCIBoot*) to address the problem.
- Theoretically, we prove that the confidence interval obtained by *ETCIBoot* is asymptotically consistent.
- The point estimates obtained by *ETCIBoot* are more accurate and robust compared with existing approaches due to the properties of bootstrap sampling, which is nicely integrated into the truth discovery procedure in *ETCIBoot*.
- Experimental results demonstrate the effectiveness of *ETCIBoot* in constructing confidence intervals as well as identifying truths.

2. PROBLEM SETTING

We first introduce terminologies and notations in this section. Then, the problem is formally defined.

DEFINITION 1. *An object is an item of interest. Its true information is defined as a truth.*

DEFINITION 2. *The reliability of a source measures the quality of its information. A source weight is proportional to its reliability, i.e., the higher the quality of a source's information, the higher its reliability, and the higher its weight.*

Assume that there are $\mathcal{S} := \{s\}_1^S$ sources, providing claims on objects $\mathcal{N} := \{n\}_1^N$, where an object may receive claims from only a subset of $\mathcal{S}$. The truths of objects $\mathcal{N}$ are denoted as $\{x_n^*\}_{n\in\mathcal{N}}$, which are unknown *a priori*. For the n-th object, $\mathcal{S}_n$ represents the set of sources which provide claims for it. The multi-source data for the n-th object is denoted as $\mathcal{X}_n := \{x_n^s\}_{s\in\mathcal{S}_n}$, where x_n^s represents the claim provided by the s-th source for the object n. The whole data collection on objects $\mathcal{N}$ is further denoted as $\mathcal{X} := \cup_{n=1}^N \mathcal{X}_n$.

For the s-th source, we assume the difference ϵ_s between its claims and truths follows a normal distribution with mean 0 and variance σ_s^2, i.e., $\epsilon_s \sim \text{Normal}(0, \sigma_s^2)$. This assumption is commonly used in existing truth discovery works [11, 12, 22]. ϵ_s captures the error of source s. As a small ϵ_s means that the claims are close to the truths, σ_s^2 measures the quality of the claims provided by the s-th source. We further denote the weight of source s as ω_s. Definition 2 implies that the larger σ_s^2, the smaller ω_s.

Truth Discovery Task. Truth discovery task is formally defined as follows: Given the multi-source data $\mathcal{X}$, the goal of a *truth discovery* approach is to obtain estimates $\hat{x}_n$ which are as close to x_n^* as possible ($\forall n \in \mathcal{N}$). Besides, for any $\alpha \in (0, 1)$, we can also provide an α-level two-sided confidence interval for each object.

We summarize the notations in Table 1.

Table 1: Notations

Notation	Definition
$\mathcal{S}$	the set of sources
$\mathcal{N}$	the set of objects
x_n^s	the claim on object n made by source s
x_n^*	the true claim of the n-th object
$\hat{x}_n$	the estimator of the claim for object n
ϵ_s	the s source's error
σ_s^2	the s-th source's variance of claims
ω_s	the weight of source s
$\mathcal{S}_n$	the subset of sources available for object n
$\mathcal{N}_s$	the subset of objects claimed by source s
$\mathcal{X}_n$	the data set available for object n
$\mathcal{X}$	the whole data set for all objects

3. METHODOLOGY

In this section, we first review some preliminaries about *truth discovery* and confidence interval in Subsection 3.1. We then introduce two main components of *ETCIBoot*: a novel strategy for data aggregation (*ETBoot*) and a method for confidence interval construction (*CIC*) in Subsections 3.2 and 3.3, respectively. The proposed *ETCIBoot* is further summarized in Subsection 3.4. Finally, we present the theoretical analysis of the confidence interval estimates obtained by *ETCIBoot* in Subsection 3.5.

3.1 Preliminary

Truth Discovery

The goal of a truth discovery task is to identify objects' truths (i.e., true information) from conflicting multi-source data. Many truth discovery methods have been proposed to estimate truths and weights iteratively. Details can be found in Section 5. We briefly introduce two iterative steps as follows.

Weight Update. Source weights play important roles in truth discovery. The underlying principle is that: If a source more often provides reliable information, it has a larger weight, and consequently this source contributes more in the truth estimation step discussed below. Based on this principle, various weight update strategies have been proposed. In this paper, we adopt the weight estimation introduced in [11]. Specifically, a source weight is inversely proportional to its total difference from the estimated truth, namely,

$$\omega_s \propto \frac{\chi^2_{(\frac{\alpha}{2},|\mathcal{N}_s|)}}{\sum_{n\in\mathcal{N}_s}(x_n^s-\hat{x}_n)^2}, \tag{1}$$

where $\chi^2_{(\frac{\alpha}{2},|\mathcal{N}_s|)}$ is the $\frac{\alpha}{2}$-th percentile of a χ^2-distribution with $|\mathcal{N}_s|$ degree. It is used to capture the effect of the number of claims so that small sources get their weights reduced.

Truth Estimation. A commonly used strategy is weighted averaging for continuous data or weighted voting for categorical data, namely,

$$\hat{x}_n = \frac{\sum_{s\in\mathcal{S}_n}\omega_s x_n^s}{\sum_{s\in\mathcal{S}_n}\omega_s}, \text{or}, \hat{x}_n = \arg\max_x \frac{\sum_{s\in\mathcal{S}_n}\omega_s \mathbb{1}(x_n^s,x)}{\sum_{s\in\mathcal{S}_n}\omega_s}, \tag{2}$$

where $\mathbb{1}(x_n^s,x)=1$ if $x_n^s=x$; otherwise it is 0. The weights are obtained at the *Weight Update* step; the truth estimated at this step will be used to update weights based on (1).

Providing proper initializations, *Weight Update* and *Truth Estimation* are iteratively executed until the convergence condition is satisfied.

Confidence Interval

Assume that an experiment has a sample set $\boldsymbol{X} = \{x_1,\cdots,x_n\}$ from $F_\mu(x)$, where F_μ is an accumulative density function (c.d.f.) with a parameter μ. An α-level confidence interval for the parameter μ is defined as follows:

DEFINITION 3. *For any $\alpha\in(0,1)$, $(\mu_{\boldsymbol{X},L},\mu_{\boldsymbol{X},R})$ is called an α-level two-sided confidence interval of a parameter μ if it satisfies the following condition:*

$$\mathbb{P}\left(\mu\in(\mu_{\boldsymbol{X},L},\mu_{\boldsymbol{X},R})\right)=\alpha. \tag{3}$$

The immediately preceding probability statement (3) can be read: Prior to the repeated independent trails of the random experiment, α is the probability that the random interval $(\mu_{\boldsymbol{X},L},\mu_{\boldsymbol{X},R})$ includes the unknown parameter μ.

Given the distribution of the experiment sample set $\boldsymbol{X}$, the exact end points of a confidence interval is defined as:

DEFINITION 4. *The exact end points of an α-level two-sided confidence interval of μ with a known c.d.f. F are:*

$$\begin{cases} \mu_{L,Exact} = \mu - \frac{\mathbb{V}\text{ar}(\mu)}{\sqrt{n}}F^{-1}(1-\alpha), \\ \mu_{R,Exact} = \mu + \frac{\mathbb{V}\text{ar}(\mu)}{\sqrt{n}}F^{-1}(\alpha); \end{cases} \tag{4}$$

where $F^{-1}(\cdot)$ is the inverse function of c.d.f. F, $\mathbb{V}\text{ar}(\mu)$ is the variance of μ, and n is the number of observed samples.

However, (4) is always unknown *a prior* because F is unknown. The major task in this paper is to construct a confidence interval estimate for each truth.

3.2 *ETBoot* Strategy

In this subsection, we introduce a novel bootstrapping-based strategy for the truth discovery task. We term this strategy as **E**stimating **T**ruth via **Boot**strapping (*ETBoot*). All existing *truth discovery* methods apply weighted averaging or voting using all sources' information. In contrast, *ETBoot* first bootstraps multiple sets of sources and then on each set of the bootstrapped sources it obtains a truth estimate based on (2). The final truth estimator is defined as the mean of these estimates. Due to the properties of bootstrapping techniques which are nicely integrated into the truth discovery procedure, *ETBoot* is more robust to the outlying claims and then achieves a better estimate of the truth. Moreover, as shown in Subsection 3.3, *ETBoot* is able to construct an α-level two-sided confidence interval of the estimated truth for any $\alpha\in(0,1)$.

The detailed procedure of *ETBoot* is as follows: for the n-th object, it obtains B estimates of its truth, i.e., $\{\hat{x}_n^b\}_{b=1}^B$, where $\hat{x}_n^b$ is obtained by the following two-step procedure:

- Step 1: *Source Bootstrap.* In this step, we randomly sample $|\mathcal{S}_n|$ sources $\mathcal{S}_n^b$ from $\mathcal{S}_n$ with replacement. The sampled data is denoted as $\boldsymbol{X}_n^b=\{x_s^n\}_{s\in\mathcal{S}_n^b}$.
- Step 2: *Truth Computation.* Based on the sampled data $\boldsymbol{X}_n^b=\{x_s^n\}_{s\in\mathcal{S}_n^b}$, $\hat{x}_n^b$ is calculated based on (2).

The final estimator $(\hat{x}_n^{Boot})$[1] for the n-th object's truth is further defined as:

$$\hat{x}_n^{Boot} = \frac{1}{B}\sum_{b=1}^B \hat{x}_n^b. \tag{5}$$

[1]We use $\cdot^{Boot}$ to represent the estimator obtained by Bootstrapping throughout the paper.

Compared with existing truth discovery methods which use (2), the proposed *ETBoot* combines results from multiple bootstrap samples instead of using all the sources at once. This enables *ETBoot* to obtain more robust estimates and confidence interval estimates as explained in Subsection 3.3.

The pseudo code of *ETBoot* for the n-th object is summarized in Algorithm 1.

Algorithm 1 *ETBoot*

Input: $\mathcal{S}_n$, $\mathcal{X}_n$, $\{\omega_s\}_{s\in\mathcal{S}_n}$, and B

1: **for** the b-th iteration ($b = 1, \cdots, B$) **do**
2: Bootstrap $\mathcal{S}_n^b$ from $\mathcal{S}_n$;
3: Extract $\boldsymbol{X}_n^b$ from $\mathcal{X}_n$;
4: Calculate $\hat{x}_n^b$ according to (2);
5: **end for**
6: Calculate $\hat{x}_n^{Boot}$ according to (5);

Output: $\hat{x}_n^{Boot}$.

3.3 Confidence Interval Construction

In this subsection, we introduce the procedure of constructing an α-level two-sided confidence interval of an object's truth. We illustrate it for the n-th object. The procedure is similar for other objects.

We denote the estimator we are interested in as $\hat{\theta}(\boldsymbol{X}_n)$ corresponding to the dataset $\boldsymbol{X}_n = \{x_n^s\}_{s\in\mathcal{S}_n}$. In our scenario, $\hat{\theta}(\boldsymbol{X}_n)$ denotes the truth estimate. For simplicity, we ignore the subscript $\cdot_n$ for $\boldsymbol{X}_n$. In a *truth discovery* task, the truth estimate is calculated as $\hat{\theta}(\boldsymbol{X}) = \frac{\sum_{s\in\mathcal{S}_n}\omega_s x_s}{\sum_{s\in\mathcal{S}_n}\omega_s}$, yielding,

$$\mathbb{E}(\hat{\theta}(\boldsymbol{X})) = x_n^*, \quad \text{and,} \quad \mathbb{V}\text{ar}(\hat{\theta}(\boldsymbol{X})) = \frac{\sum_{s\in\mathcal{S}_n}\omega_s^2\sigma_s^2}{(\sum_{s\in\mathcal{S}_n}\omega_s)^2}. \quad (6)$$

The corresponding estimate of $\mathbb{V}\text{ar}(\hat{\theta}(\boldsymbol{X}))$ is defined as $\widehat{\mathbb{V}\text{ar}}(\hat{\theta}(\boldsymbol{X})) = \frac{\sum_{s\in\mathcal{S}_n}\omega_s^2\hat{\sigma}_s^2}{(\sum_{s\in\mathcal{S}_n}\omega_s)^2}$ where $\hat{\sigma}_s^2 = \frac{\sum_{n\in\mathcal{N}_s}(x_n^s - \hat{x}_n^{Boot})^2}{N_s - 1}$ and $\hat{x}_n^{Boot}$ is obtained by *ETBoot*. To obtain a confidence interval of the truth x_n^*, we first construct a statistic T which is related to x_n^*, and then estimate the accumulated density function of $T \sim F(t)$. In our scenario, T is defined as follows:

$$T = \frac{\hat{\theta}(\boldsymbol{X}) - x_n^*}{[\widehat{\mathbb{V}\text{ar}}(\hat{\theta}(\boldsymbol{X}))]^{\frac{1}{2}}/\sqrt{|\mathcal{S}_n|}}, \quad (7)$$

which measures the error between $\hat{\theta}(\boldsymbol{X})$ and x_n^*. The confidence interval of x_n^* is available once the distribution of T is determined. More precisely, let $T^{(\alpha)}$ indicate the $(100\cdot\alpha)$-th percentile of T, i.e., $\alpha = \int_{-\infty}^{T^{(\alpha)}} dF(t)$. Thus, we have that

$$\mathbb{P}\left(T^{(\alpha/2)} \le \frac{\hat{\theta}(\boldsymbol{X}) - x_n^*}{[\widehat{\mathbb{V}\text{ar}}(\hat{\theta}(\boldsymbol{X}))]^{\frac{1}{2}}/\sqrt{|\mathcal{S}_n|}} \le T^{(1-\alpha/2)}\right) = \alpha. \quad (8)$$

Moreover, an α-level two-sided confidence interval of x_n^* is naturally implied in (8), that is,

$$\left(\hat{\theta}(\boldsymbol{X}) - \frac{T^{(1-\alpha/2)}[\widehat{\mathbb{V}\text{ar}}(\hat{\theta}(\boldsymbol{X}))]^{\frac{1}{2}}}{\sqrt{|\mathcal{S}_n|}}, \quad (9)$$

$$\hat{\theta}(\boldsymbol{X}) - \frac{T^{(\alpha/2)}[\widehat{\mathbb{V}\text{ar}}(\hat{\theta}(\boldsymbol{X}))]^{\frac{1}{2}}}{\sqrt{|\mathcal{S}_n|}}\right). \quad (10)$$

Thus, the width of the confidence interval is proportional to $\frac{1}{\sqrt{|\mathcal{S}_n|}}$. It implies that if an object is claimed by more sources then the width of its truth's confidence level is smaller, and vice versa. Especially, when the long-tail multi-source data is involved, this phenomenon is clearer.

However, as the T-percentile is usually unknown *a priori*, estimation of $T^{(\alpha)}$ is required. One commonly used strategy is bootstrap sampling [1, 6, 8, 10, 14]. Note that at the b-th iteration of *ETBoot* (Algorithm 1), we have bootstrapped $\boldsymbol{X}_n^b$. Based on $\boldsymbol{X}_n^b$, we are able to calculate both $\hat{\theta}(\boldsymbol{X}_n^b)$ and $\widehat{\mathbb{V}\text{ar}}(\boldsymbol{X}_n^b)$, yielding an estimator $\widehat{T}_b$ for the statistic T, that is,

$$\widehat{T}_b = \frac{\hat{\theta}(\boldsymbol{X}_n^b) - \hat{\theta}(\boldsymbol{X})}{[\widehat{\mathbb{V}\text{ar}}(\hat{\theta}(\boldsymbol{X}_n^b))]^{\frac{1}{2}}/\sqrt{|\mathcal{S}_n|}}. \quad (11)$$

Moreover, the estimate of $T^{(\alpha)}$ is defined as follows:

$$\widehat{T}^{(\alpha)} = \sup\left\{t \in \{\widehat{T}_1, \cdots, \widehat{T}_B\} : \frac{\#(\widehat{T}_b \le t)}{B} \le \alpha\right\}. \quad (12)$$

(12) provides estimates of (9) and (10). Thus, the estimate of an α-level two-sided confidence interval is defined as follows:

$$\left(\hat{\theta}(\boldsymbol{X}) - \frac{\widehat{T}^{(1-\alpha/2)}[\widehat{\mathbb{V}\text{ar}}(\hat{\theta}(\boldsymbol{X}))]^{\frac{1}{2}}}{\sqrt{|\mathcal{S}_n|}}, \quad (13)$$

$$\hat{\theta}(\boldsymbol{X}) - \frac{\widehat{T}^{(\alpha/2)}[\widehat{\mathbb{V}\text{ar}}(\hat{\theta}(\boldsymbol{X}))]^{\frac{1}{2}}}{\sqrt{|\mathcal{S}_n|}}\right). \quad (14)$$

We summarize the procedure of constructing confidence intervals as *CIC*, i.e., **C**onfidence **I**nterval **C**onstruction. Its pseudo is presented in Algorithm 2 for the n-th object.

Algorithm 2 *CIC*

Input: $\{\boldsymbol{X}_n^b\}_{b=1}^B$, $\hat{x}_n^{Boot}$, and α.

1: Calculate $\hat{\sigma}_s^2$ for $s \in \mathcal{S}_n$;
2: **for** the iteration b ($b = 1, \cdots, B$) **do**
3: Calculate $\widehat{\mathbb{V}\text{ar}}(\hat{\theta}(\boldsymbol{X}_n^b))$ and $\widehat{T}_b$ according to (11);
4: **end for**
5: Choose $\widehat{T}(1-\alpha/2)$ and $\widehat{T}(\alpha/2)$ according to (12);

Output: Endpoints calculated based on (13) and (14).

3.4 *ETCIBoot* Algorithm

So far, we introduce the update for source weights (i.e., (1)), a new truth estimation strategy, *ETBoot*, and the construction of confidence intervals for truths via *CIC*. Combining them together, we propose a novel truth discovery approach, **E**stimating **T**ruth and **C**onfidence **I**nterval via **Boot**strapping (*ETCIBoot*), to automatically construct confidence intervals as well as identify objects' truths. The main component of the proposed *ETCIBoot* consists of the following three steps:

(i) *Weight Update.* Given initialization of truth $\{x_n^0\}$, source weights are updated based on (1).

(ii) *Truth Estimation.* With source weights computed from previous step, for each object n, we obtain truth estimators via *ETBoot* at this step to obtain $\hat{x}_n^{Boot}$ associated with $\{\boldsymbol{X}_n^b\}_{b=1}^B$.

(iii) *Confidence Interval Construction.* For all objects, the estimation of confidence intervals for their truths are obtained via *CIC*.

The above two steps are executed iteratively until no truth estimates change anymore. The pseudo code of the proposed *ETCIBoot* algorithm is shown in Algorithm 3.

Algorithm 3 *ETCIBoot*

Input: the whole data collection $\mathcal{X}$, confidence level α, and the number of bootstrapping samples B.

```
Initialize truths x_1^{*,0}, ..., x_N^{*,0} as average;
while the convergence condition is not satisfied do
    Compute ω_s for each source s according to (1);
    for each object n (n = 1, ..., N) do
        Conduct ETBoot to obtain x̂_n^{Boot};
        Calculate the confidence interval CI_n(α) via CIC;
    end for
end while
```

Output: $\{\hat{x}_n^{Boot}\}_1^N$ and confidence interval $\{CI_n(\alpha)\}_1^N$.

3.5 Theoretical Analysis

In this subsection, we present the theoretical analysis on the confidence interval estimates, i.e., (13) and (14), obtained via *ETCIBoot*. We first prove that $\widehat{T}$ converges to T in distribution and present it in Proposition 1.

PROPOSITION 1. *Assume that $x_n^s \sim \mathcal{N}(x_n^*, \sigma_s^2)$, for any $s \in \mathcal{S}_n$. Let T and T^* be defined as (7) and (11), respectively. Then, we have that*

$$\lim_{|\mathcal{S}_n| \to \infty} \| \mathbb{P}^*(\widehat{T} \leq t) - \mathbb{P}(T \leq t) \| = 0, \quad a.s., \tag{15}$$

where $\mathbb{P}^*$ is the probability calculated based on the bootstrapping sample distribution, $|\mathcal{S}_n|$ is the Cardinality of $\mathcal{S}_n$, t is any real number, and a.s. means 'almost surely'.

PROOF. See Appendix A for a detailed proof. □

Proposition 1 is a straightforward result from Theorem 1 in [14], where the author provides sufficient conditions to guarantee the convergence of the bootstrapping samples. Thus, the proof of Proposition 1 is to testify whether the *ETCIBoot* satisfies these sufficient conditions, as shown in Appendix A. Proposition 1 shows that the bootstrapping estimator $\widehat{T}$ converges to T in distribution. It enables us to use the bootstrapping distribution to approximate the unknown distribution F for confidence interval construction.

Next, in Proposition 2, we show that the upper end point of an α-level one-sided confidence interval obtained via *ETCIBoot* is close to that from the theoretical distribution.

PROPOSITION 2. *Given $T \sim F(x)$, $\widehat{T} \sim \widehat{F}(x)$ and a dataset $\boldsymbol{X}$, we have that*

$$\hat{\theta}_{\widehat{T},\boldsymbol{X}}(\alpha) = \hat{\theta}_{T,\boldsymbol{X}}(\alpha) + O_p(n^{-3/2}), \tag{16}$$

where $\mathbb{P}^*(\theta(\boldsymbol{X}) \leq \hat{\theta}_{\widehat{T},\boldsymbol{X}}(\alpha)) = \alpha$, $\mathbb{P}(\theta(\boldsymbol{X}) \leq \hat{\theta}_{T,\boldsymbol{X}}(\alpha)) = \alpha$, $n = |\boldsymbol{X}|$, and O_p means the order holds in probability.

PROOF. See Appendix B for a detailed proof. □

Proposition 2 shows that the endpoint of an α-level one-sided confidence interval obtained by bootstrapping $\widehat{T}$ is close to that obtained by T, provided that there are enough samples. As any α-level two-sided confidence interval can be obtained by two one-sided confidence intervals, the results ((16)) also hold for (13) and (14). In truth discovery tasks, *ETCIBoot* is able to provide more accurate confidence intervals for the objects' truths, if they receive more claims. This result is more obvious especially on long-tail data.

4. EXPERIMENTS

In this section, we evaluate the proposed *ETCIBoot* method on both simulated and real-world datasets. We first introduce the experimental setup in Subsection 4.1. Then, we test the *ETCIBoot* and baselines on simulated datasets generated in different scenarios and real-world datasets in Subsections 4.2 and 4.3, respectively. Experimental results show that: (1) *ETCIBoot* outperforms the state-of-the-art truth discovery methods in most cases, and (2) *ETCIBoot* can provide accurate confidence interval estimates.

4.1 Experimental Setup

In this part, we introduce the baseline methods and discuss the measurements for evaluation.

Baseline Methods. For all truth discovery methods, we conduct them on the same input data in an unsupervised manner. Although ground truths are available, we only use them for evaluation. For different data types, different baselines are adopted, including both the naive conflict resolution methods and the state-of-the-art truth discovery methods. More precisely, for continuous data we use Median, Mean, CATD [11], CRH [12] and GTM [22]. Baselines used for categorical data include: Voting, Accusim [5], 3-estimate [7], CRH [12], Investment [18], CATD [11], ZenCrowd [3], Dawid&Skene [2], and TruthFinder [21]. Details of baselines are discussed in the related work (Section 5).

Measurements. As the experiments involve both continuous and categorical data, we introduce different measurements. For data of continuous type, we adopt both the mean of absolute error (*MAE*) and the root of mean square error (*RMSE*); *Error Rate* is used for date of categorical type. The details of the measurements are:

- *MAE*: *MAE* measures the L^1-norm between the methods' output and the ground truths. It tends to penalize more on small errors.
- *RMSE*: *RMSE* measures the L^2-norm between the methods' output and the ground truths. It tends to penalize more on the large distance and less on the small distance comparing with *MAE*.
- *Error Rate*: *Error Rate* is defined as the percentage of mismatched values between the output of each method and the ground truths.

Note That: the smaller the measurement value, the closer to ground truths the methods' output. Therefore, for all measurements, the smaller the value, the better the method.

4.2 Simulated Datasets

In this subsection, we test the proposed *ETCIBoot* on several simulated datasets, which capture different scenarios involving various distributions of source reliability. We first introduce the procedure of generating simulated datasets, and then test the effectiveness of *ETCIBoot* in identifying truths comparing with baselines on these datasets. Last but not least, we compare the confidence intervals obtained by *ETCIBoot* with that by theoretical distribution and show the advantage of bootstrapping.

Table 2: Comparison on simulated data: all scenarios

Method	Scenario 1 (Uniform(0, 1))		Scenario 2 (Gamma(1, 3))		Scenario 3 (FoldedNormal(1, 2))		Scenario 4 (Beta(1, $\frac{1}{2}$))	
	MAE (10^{-3})	*RMSE* (10^{-1})	*MAE* (10^{-3})	*RMSE* (10^{-1})	*MAE* (10^{-3})	*RMSE* (10^{-1})	*MAE* (10^{-3})	*RMSE* (10^{-1})
ETCIBoot	**1.200**	**2.724**	**0.700**	**1.599**	**0.400**	**0.973**	**0.600**	**.1351**
CATD	1.300	2.903	0.800	1.708	0.500	1.043	**0.600**	0.141
CRH	6.300	13.560	6.900	14.781	1.700	3.729	4.500	0.968
Median	4.000	9.275	2.800	6.648	1.300	2.914	2.400	0.551
Mean	8.000	17.224	11.700	24.988	2.000	4.369	6.400	1.362
GTM	4.000	8.549	4.000	8.593	1.200	2.597	2.700	0.583

Data Generation. The procedure of generating simulated data is shown as follows:

(i) We first generate a vector of the number of claims $\boldsymbol{C}$, e.g., $\boldsymbol{C} = (5, 10, 15, \cdots, 50)$.

(ii) For each $c_i \in \boldsymbol{C}$, there are $o_i = e^7 \cdot c_i^{-1.5}$ objects which will receive c_i claims. This power law function is used to create the long-tail multi-source data. Thus, there are totally $O = \sum_i o_i$ objects and $S = \max_i\{c_i\}$ sources.

(iii) For each source, we randomly generate its reliability $\sigma_s^2 \sim F$, where F is a pre-defined distribution. Thus, for each source, its claims are generated from Normal$(0, \sigma_s^2)$. Here, σ_s^2 captures reliability degree of the s-th source's information. The larger value the σ_s^2, the lower reliability degree of the s-th source.

Experiments. In the following experiments, we simulate different scenarios via different source reliability distributions F. We set $\boldsymbol{C} = 70 : 100$; thus, there are 31 objects and 100 sources. Note that the number of objects is not large. This is used to better display the experimental results on the confidence interval estimates. To reduce the randomness, we repeat the experiment 100 times and report the average results. As the simulated data is continuous, *MAE* and *RMSE* are used for evaluation. We simulate 4 scenarios and the detail of each scenario is discussed as follows. Note that σ_s^2 represents the source reliability degree. The larger value the σ_s^2, the lower reliability degree the source.

Scenario 1: $\sigma_s^2 \sim$ Uniform$(0, 1)$. In this scenario, all source reliability degrees are uniformly distributed in $(0, 5)$.

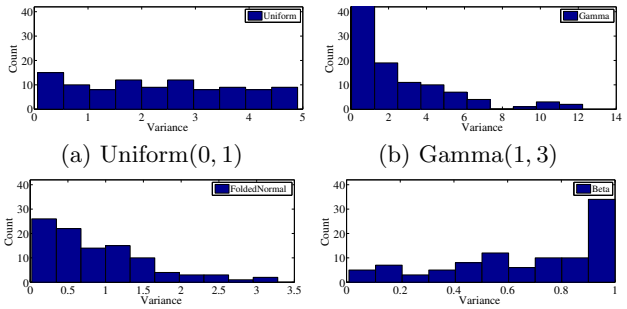
Scenario 2: $\sigma_s^2 \sim$ Gamma$(1, 3)$. In this scenario, most of the sources are reliable with high reliability degrees. However, there are a few unreliable sources with very small reliability degrees.

Scenario 3: $\sigma_s^2 \sim$ FoldedNormal$(1, 2)$. As Folded Normal is a long-tail distribution, in this scenarios, it generates a few unreliable sources. Compared with Scenarios 1 and 2, the reliable sources have higher reliability degrees.

Scenario 4: $\sigma_s^2 \sim$ Beta$(1, \frac{1}{2})$. In this scenario, source reliability degrees are within $0 \sim 1$. Compared with other scenarios, there are much more reliable sources.

We show the histograms of the source variances in Figure 2, which implies that the simulated data covers various scenarios with varying source reliability distributions. We report the results in terms of *MAE* and *RMSE* in Table 2.

Comparison with Baselines. Table 2 shows that the proposed *ETCIBoot* outperforms all baselines in all scenarios in terms of both *MAE* and *RMSE*. When estimating the truth for each object n, *ETCIBoot* obtains multiple truth estimates which are calculated according to (2) based on the bootstrapped claims. Then, the final truth estimator is defined as the average of these estimates. Experimentally, we generate $10 * |\mathcal{S}_n|$ bootstrapping samples. Due to the properties of bootstrapping, *ETCIBoot* is robust to the outlying claims provided by some sources. However, as existing truth discovery methods typically compute weighted mean to obtain one single point estimate, they are more sensitive to the outlying claims. So, the *ETCIBoot* performs better than baselines as confirmed in the experimental results. Moreover, as there are more reliable sources in Scenarios 3 and 4, the results are better compared with those in Scenarios 1 and 2. It confirms the underlying intuition of truth discovery: the more the reliable sources, the better the results.



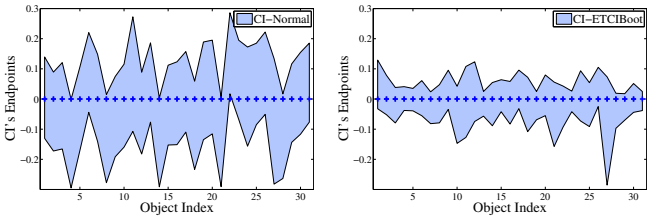
(a) Uniform$(0, 1)$ (b) Gamma$(1, 3)$

(c) FoldedNormal$(1, 2)$ (d) Beta$(1, \frac{1}{2})$

Figure 2: ***Histograms of source variances***

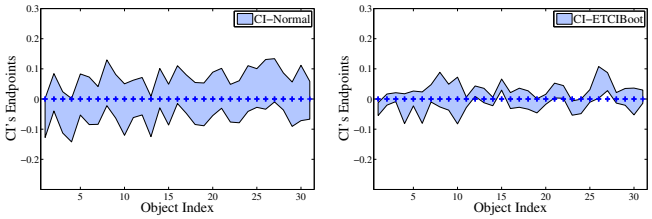
Confidence Interval Comparison. For confidence interval comparison, we compare the results of *ETCIBoot* with that obtained by theoretical distribution, i.e., normal distribution. Note that $\hat{x}_n \sim$ Normal$(x_n^*, \frac{\sum_{s\in\mathcal{S}_n} \omega_s^2\sigma_s^2}{(\sum_{s\in\mathcal{S}_n} \omega_s)^2})$ (based on (2)). As the true σ_s^2 is known for each source, we know the theoretical distribution for $\hat{x}_n$, based on which we can further obtain the 95%-level confidence interval. We term the confidence interval obtained in this way as *CI-Normal*. The confidence interval (i.e., (13) and (14)) for the truths' estimators, which is obtained by the *ETCIBoot* using the bootstrapping technique, is referred to as *CI-ETCIBoot*.

We report the results in Scenarios $1 \sim 4$ in Figures $3 \sim 6$, respectively. From Figures $3 \sim 6$, we can draw the following conclusions: (1) The *CI-ETCIBoot* is much smaller than *CI-Normal* in all simulated scenarios. Note that the smaller the confidence interval, the more confident the estimator. For example, in Scenario 1 the shaded area (i.e., the area between the lower and upper bound curves) of *CI-Normal* in Figure 3(a) is larger than that of *CI-ETCIBoot* in Figure 3(b). Similar conclusions can be drawn in other



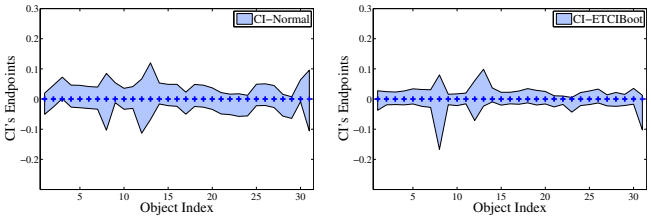
(a) *CI-Normal* (b) *CI-ETCIBoot*

Figure 3: *Scenario* 1*:* Uniform$(0, 5)$



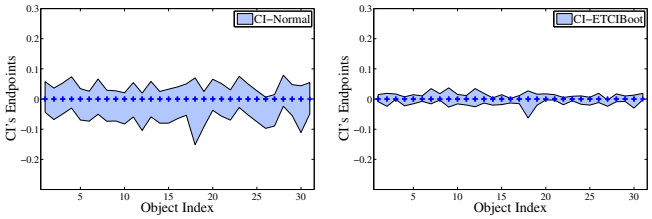
(a) *CI-Normal* (b) *CI-ETCIBoot*

Figure 4: *Scenario* 2*:* Gamma$(1, 3)$



(a) *CI-Normal* (b) *CI-ETCIBoot*

Figure 5: *Scenario* 3*:* FoldedNormal$(1, 2)$

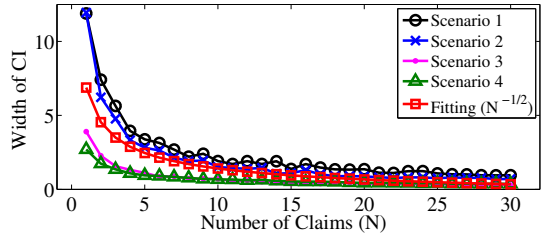


(a) *CI-Normal* (b) *CI-ETCIBoot*

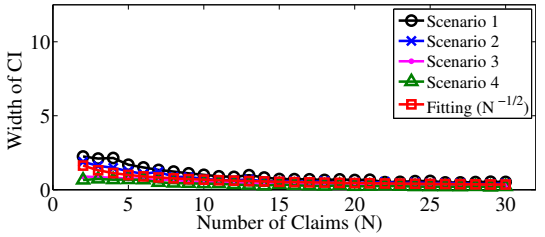
Figure 6: *Scenario* 4*:* Beta$(1, .5)$

scenarios. Thus, the experimental results show the power of the *ETCIBoot* on constructing effective confidence intervals. (2) As most sources are reliable in Scenarios $2 \sim 4$, comparing with Scenario 1, the width of *CI-ETCIBoot* or *CI-Normal* in other scenarios is smaller, which indicates the higher overall confidence in these scenarios.

Next we conduct experiments to illustrate the relationship between the width of confidence interval and the number of claims on long-tail data. We follow the same procedure to generate the simulated data, except that we choose the number of claims as 2 to 30. If there is only one claim, it is impossible to construct the confidence interval. We present the width of *CI-Normal* and *CI-ETCIBoot* in all scenarios in Figures 7(a) and 7(b), respectively. Meanwhile, we also fit them into a polynomial function of N ($N^{-\frac{1}{2}}$), respectively. The red line with square marker represents the fitting line, averaging over all scenarios. From Figure 7, we can see that the width of the 95% confidence interval, obtained via either normal distribution or *ETCIBoot*, decreases with respect to the number of claims at an error rate $N^{-\frac{1}{2}}$, where N is the number of claims. It confirms the theoretical analysis that if an object receives more claims then its estimator is more accurate. Moreover, the width of *CI-ETCIBoot* is much smaller than that of *CI-Normal*, which demonstrates that *ETCIBoot* is able to provide a more confident estimator. This advantage is achieved by incorporating bootstrapping techniques into truth discovery procedure in *ETCIBoot*.



(a) *CI-Normal*



(b) *CI-ETCIBoot*

Figure 7: *Simulated data in all scenarios: Confidence Interval width w.r.t. the number of claims* (N)

4.3 Real-World Datasets

In this subsection, we present the experimental results on two continuous datasets and two categorical datasets. Experiments show that the proposed *ETCIBoot* is able to obtain more accurate estimates of truths comparing with baselines. We first introduce the description of the datasets and then report the results.

Continuous Data

Dataset Description. The following datasets of continuous data type are used in experiments:

- Indoor Floorplan Dataset: We develop an Android App that can estimate the walking distances of smartphone users though multiplying their step sizes by step count inferred using the in-phone accelerometer. There are totally 247 users and 129 objects (i.e., indoor hallways). The ground truth of the hallway length is obtained by manually measuring the indoor hallways. The goal is to estimate the distance of indoor hallways from the data provided by a crowd of users.

- Flight Status Dataset: The flight data [13] is collected by extracting departure/arrival information for $11,512$ flights from 38 sources on every day in December 2011. We present the time in terms of the minutes from 00:00. There are $11,146$ flights that have departure/arrival ground truths. The goal is to estimate the departure/arrive time for each flight.

Results Analysis. We present the results of *ETCIBoot* and baselines with respect to *MAE* and *RMSE* on the continuous datasets in Table 3. The experimental results show that the proposed *ETCIBoot* can achieve the best performance on both datasets.

Table 3: Comparison on continuous data

Method	Indoor Floorplan		Flight Status	
	MAE (10^0)	*RMSE* (10^1)	*MAE* (10^0)	*RMSE* (10^3)
ETCIBoot	**.9219**	**1.2992**	**.0310**	**.9933**
CATD	0.9960	1.3845	1.077	8.120
CRH	1.1929	1.5955	1.074	8.094
Median	1.3797	1.7860	1.070	8.020
Mean	1.7851	2.2846	1.055	7.893
GTM	1.2845	1.6823	1.078	8.132

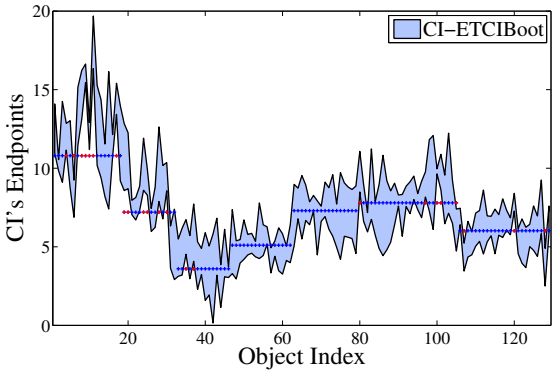


Figure 8: *Indoor Floorplan dataset: CI-ETCIBoot*

On Indoor Floorplan dataset, as the number of objects is small, we also present the confidence intervals obtained by *ETCIBoot* for each object in Figure 8. The figure shows that in most cases the confidence intervals provided by *ETCIBoot* contains the corresponding objects' truths. However, there are some confidence intervals which do not contain truths. A possible reason is: These objects are claimed by a few sources and the information provided by these sources is far away from the truth. Take the 9-th object for example. There are only 4 sources which provide claims, among which the smallest value is 14.3 that is still very larger than the ground truth 10.8. As a result, it is impossible to correctly identify these objects' truths for any truth discovery method. Therefore, the confidence interval estimates obtained by *ETCIBoot* do not contain the truths for these objects.

On Flight Status dataset, the data on each day is treated as a single data collection. As there are many flights only claimed by a few sources, the performance of baselines is not satisfactory. We conduct a case study on Day 1 dataset. We count the statistics on how many claims of an object receives to show the long-tail phenomenon: (1) there are about 61.1% of flights which only receives claims from at most 5 out of 38 sources; (2) only 2.3% of flights have received claims from more than 25 sources. Similar phenomenon can be found on other days' data. Consequently, we can see that the proposed *ETCIBoot* outperforms all baselines, as shown in Figure 9. We do not present the confidence interval for the flights due to the page limit and the large number of flights.

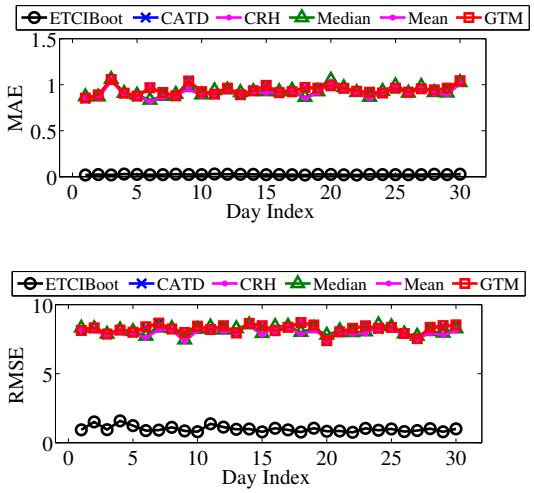


Figure 9: *Comparison on Flight data over 30 days*

Categorical Data

Dataset Description. We introduce the details of two categorical datasets and their tasks as follows:

- SFV Dataset: SFV dataset is built upon the annal Slot Filling Validation (SFV) competition of the NITS Text Analysis Conference Knowledge Base Population track [9]. In this task, given a query (an object), e.g., the birthday of Obama, 18 slot filling systems (sources) extract useful claims independently from a large-scale corpus. The 2011 SFV dataset[2] contains 2,538 claims from 18 sources for 328 objects. The goal is to extract the true answer for each query from the systems' claims.

- Game Dataset: Game dataset [11] collects answers from multiple users based on a TV game show "Who Wants to Be a Millionaire" via an Android App. There are 37,029 Android users and 2,103 questions. Ground truths are available for evaluation. The goal is to identify each question's answer from the users' answers.

Results Analysis. For categorical data, we first encode the claims into probability vectors and then apply the methods proposed for continuous data, such as *ETCIBoot*, CATD, etc. The detailed procedure is: For a question with 4 possible choices, the first choice is encoded into a 4-element vector $(1, 0, 0, 0)$. In Tables 4 and 5, we present the experimental results of the proposed *ETCIBoot* as well as baselines on the SFV and Game datasets, respectively.

On SFV dataset, there are only 18 sources, so we have a limited number of sources to bootstrap at each iteration of *ETCIBoot*. Thus, the result of the proposed *ETCIBoot* (.0945) is not the best, but still comparable with the two best methods: AccuSim (.0701) and TruthFinder (.0793).

On Game dataset, the number of sources (37,029) is sufficient for bootstrapping. Although CATD performs best among all baselines, the proposed *ETCIBoot* achieves even better performance compared with CATD. Especially, on the Levels 8, 9, and 10, the proposed *ETCIBoot* improves the results by 33.28%, 50.00% and 33.30%, respectively, when

[2]http://www.nist.gov/tac/2011/

Table 5: 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)	All Levels (2103)
ETCIBoot	.0165	**.0271**	**.0241**	**.0217**	**.0395**	**.0505**	**.0481**	**.0870**	**.0707**	**.1364**	**.0385**
CATD	**.0132**	**.0271**	.0276	.0290	.0435	.0596	**.0481**	.1304	.1414	.2045	.0485
CRH	.0264	**.0271**	.0345	.0435	.0593	.0872	.0856	.2609	.3535	.4545	.0866
ZenCrowd	.0330	.0305	.0345	.0471	.0593	.0872	.0856	.2754	.3636	.5227	.0899
AccuSim	.0264	.0305	.0345	.0507	.0632	.0963	.0909	.2826	.3636	.5000	.0913
3-Estimates	.0264	.0305	.0310	.0507	.0672	.1055	.0963	.2971	.3737	.5000	.0942
Dawid&Skene	.0297	.0305	.0483	.0507	.0672	.1101	.0963	.2971	.3636	.5227	.0975
Voting	.0297	.0305	.0414	.0507	.0672	.1101	.1016	.3043	.3737	.5227	.0980
Investment	.0330	.0407	.0586	.0761	.0870	.1239	.1283	.3406	.3838	.5455	.1151
TruthFinder	.0693	.0915	.1241	.0942	.1581	.2294	.2674	.3913	.5455	.5455	.1816

Table 4: Comparison on *SFV* dataset

Method	Error Rate
ETCIBoot	.0945
CATD	.1037
CRH	.0854
ZenCrowd	.1010
AccuSim	**.0701**
3-Estimates	.1128
Voting	.1128
Dawid&Skene	.0985
Investment	.2896
TruthFinder	.0793

compared with the best baseline CATD. As *ETCIBoot* integrates bootstrapping techniques into the truth discovery procedure, it is more robust to the wrong claims compared with baselines. Thus, *ETCIBoot* can obtain better results as the experiments show. Note that there are 81 objects on which no sources provide correct answers. Therefore, the lowest error rate for any truth discovery method is .0380. *ETCIBoot* can achieve error rate at .0385, which shows its effectiveness in identifying truths.

5. RELATED WORK

Truth discovery has become an eye-catching term recently and many truth discovery methods have been proposed to identify true information (i.e., truths) from the conflicting multi-source data. The advantage of truth discovery over the naive aggregation methods such as averaging or voting is that it can capture the variance in sources' reliability degrees. Therefore, truth discovery methods can estimate source reliability automatically from the data, which is integrated into truth computation as source weight. Consequently, the more reliable sources contribute more in the final aggregation step.

A large variety of truth discovery methods have been designed to jointly estimate truths and source reliability. In [12], the authors formulate the truth discovery task into an optimization framework (CRH). They propose to minimize the overall weighted distance between claims from sources and aggregated results. CATD [11] is a statistical method that has been proposed to deal with long-tail phenomenon in truth discovery tasks, where confidence interval is incorporated in source weight estimation. However, CATD does not consider the long-tail phenomenon on objects, which can be solved by *ETCIBoot*. In [22], the authors propose a probabilistic model based truth discovery framework (GTM). Both AccuSim [5] and TruthFinder [21] adopt Bayesian analysis to estimate source reliability and update truths iteratively. In [18], the authors take the prior knowledge on truth and background information into consideration and propose a truth discovery method Investment. In [7], 3-Estimate considers the difficulty of getting the truth for each object when calculating source weights as well as complement vote. Dawid&Skene [2] and ZenCrowd [3] propose to use Expectation-Maximization technique to update source weights and truths simultaneously, based on a confusion matrix.

However, most existing truth discovery methods have the following limitations: (1) As most of them apply weighted averaging, they are sensitive to outlying claims, and (2) they focus on point estimation of the truth, where important confidence information is missing. To the best of our knowledge, this is the first paper to illustrate the importance of confidence interval estimation in *truth discovery*, and proposes an effective method (*ETCIBoot*) to address it. By integrating bootstrapping into truth discovery, *ETCIBoot* is robust compared with the state-of-the-art truth discovery methods.

6. CONCLUSIONS

In this paper, we first illustrate the importance of confidence interval estimation in truth discovery, which has never been discussed in existing work. To address the problem, we propose a novel truth discovery method (*ETCIBoot*) to construct confidence interval estimates as well as identify truths. The bootstrapping techniques are nicely integrated into the truth discovery procedure in *ETCIBoot*. Due to the properties of bootstrapping, the estimators obtained by *ETCIBoot* are more accurate and robust compared with the state-of-the-art truth discovery approaches. Theoretically, we prove that the confidence interval obtained by *ETCIBoot* is asymptotically consistent. Experimentally, we demonstrate that *ETCIBoot* is not only effective in constructing confidence intervals but also able to obtain better truth estimates.

7. ACKNOWLEDGEMENTS

This work was sponsored in part by US National Science Foundation under grant IIS-1319973, IIS-1553411 and CNS-1566374. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agency.

8. REFERENCES

[1] D. Cheng and Y. Liu. Parallel gibbs sampling for hierarchical dirichlet processes via gamma processes equivalence. In *Proc. of KDD*, pages 562–571, 2014.

[2] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Appl. Stat.*, pages 20–28, 1979.

[3] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proc. of WWW*, pages 469–478, 2012.

[4] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proc. of KDD*, pages 601–610, 2014.

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

[6] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. CRC press, 1994.

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

[8] R. V. Hogg and A. T. Craig. *Introduction to Mathematical Statistics*. Wiley, 1978.

[9] H. Ji, R. Grishman, H. T. Dang, K. Griffitt, and J. Ellis. Overview of the tac 2010 knowledge base population track. In *Proc. of TAC*, 2010.

[10] A. Q. Li, A. Ahmed, S. Ravi, and A. J. Smola. Reducing the sampling complexity of topic models. In *Proc. of KDD*, pages 891–900, 2014.

[11] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han. A confidence-aware approach for truth discovery on long-tail data. *PVLDB*, 8(4):425–436, 2014.

[12] 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.

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

[14] R. Y. Liu. Bootstrap procedures under some non-i.i.d. models. *Ann. Stat.*, 16(4):1696–1708, 1988.

[15] F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In *Proc. of KDD*, pages 745–754, 2015.

[16] A. Marian and M. Wu. Corroborating information from web sources. *Data Eng. Bull.*, pages 11–17, 2011.

[17] C. Meng, W. Jiang, Y. Li, J. Gao, L. Su, H. Ding, and Y. Cheng. Truth discovery on crowd sensing of correlated entities. In *Proc. of SenSys*, pages 169–182, 2015.

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

[19] 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.

[20] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher. On truth discovery in social sensing: A maximum likelihood estimation approach. In *Proc. of IPSN*, pages 233–244, 2012.

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

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

[23] B. Zhao, B. I. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *PVLDB*, pages 550–561, 2012.

[24] Z. Zhao, J. Cheng, and W. Ng. Truth discovery in data streams: A single-pass probabilistic approach. In *Proc. of CIKM*, pages 1589–1598, 2014.

[25] S. Zhi, B. Zhao, W. Tong, J. Gao, D. Yu, H. Ji, and J. Han. Modeling truth existence in truth discovery. In *Proc. of KDD*, pages 1543–1552, 2015.

APPENDIX

A. PROOF OF PROPOSITION 1

For the object n, we have $\mathcal{S}_n$ with $|\mathcal{S}_n| = n$. Denote the distribution of a sample x_n^i as $G_i(\cdot)$. Based on the assumption, $x_n^i \sim \text{Normal}(\hat{x}_n^*, \sigma_i^2) = G_i(x_n^i)$. As shown in [14], to prove Proposition 1 we only need to prove the following conditions:

(a) There exists a non-lattice distribution H with mean 0 and variance 1, and a sequence k_n with $\frac{k_n}{\log n} \to \infty$, such that k_n of the population $G_i's$ are of the form $G_i(x) = H(\frac{x-\mu_i}{\sigma_i})$ with σ_i's bounded away from 0;

(b) $\mathbb{E}\left(|X_i|^{3+\delta_0}\right) \leq M_1 < \infty$ for some $\delta_0 > 0$;

(c) $\liminf_{n\to\infty} v_n^2 > 0$ and $\frac{1}{n}\sum_{i=1}^n (\mu_i - \bar{\mu}_n)^2 = o(n^{-\frac{1}{2}})$;

(d) H is continuous; $(\exists \delta > 0)\ \mathbb{E}\left(|X_i|^{6+\delta}\right) \leq M_2 < \infty$,

where $\mu_i = \hat{x}^*$, and $\bar{\mu} = \frac{1}{n}\sum_{i\in\mathcal{S}_n} \mu_i$. Namely, $\forall i, \bar{\mu} = \hat{x}^* = \mu_i$. Next, we prove the sufficient conditions point by point.

Proof of (a). As introduced in Section 2, $x_s \sim \text{Normal}(x^*, \sigma_s^2)$, where $\sigma_s^2 > 0$. Let H be the standard normal distribution, i.e., Normal$(0, 1)$. As any continuous distribution is non-lattice, H is a non-lattice distribution. Moreover, let $k_n = n$ and $G_i = H(\frac{x-\mu_i}{\sigma_i})$. We have $\frac{n}{\log n} \to \infty$.

Proof of (b). For the normal distribution, we have that

$$\mathbb{E}\left(|X_i|^p\right) = \sigma_i^p \frac{2^{\frac{p}{2}}\Gamma(\frac{p+1}{2})}{\sqrt{\pi}}, \tag{17}$$

where $\Gamma(\cdot)$ is the gamma function, i.e., $\Gamma(n+1) = n\Gamma(n)$. Let $\delta_0 = 1$, we have that $\mathbb{E}\left(|X_i|^4\right) = \sigma_i^4 \frac{4\Gamma(\frac{5}{2})}{\sqrt{\pi}} = 3\sigma_i^4 \triangleq M_1 < \infty$.

Proof of (c). (i)$\forall s$, $\sigma_s^2 > 0$ and $v_n^2 = \frac{1}{n}\sum_{i=1}^n \sigma_n^2 > 0$. (ii) $\forall i$, $\mu_i = \bar{\mu}_n$.

Proof of (d). As shown in the proof of (a), H is a normal distribution which is continuous. Let $\delta = 1$ and combine with (17), yielding that $\mathbb{E}\left(|X_i|^8\right) = \sigma_i^8 \frac{2^{\frac{8}{2}}\Gamma(\frac{9}{2})}{\sqrt{\pi}} = 105\sigma_i^8 < \infty$. As shown above, all the conditions are satisfied. Thus,

$$\begin{aligned} &\mathbb{P}(T \leq t) = \Phi(t) + \tfrac{\bar{\mu}_{3,n}}{6v_n^3\sqrt{n}}(2t^2+1)\phi(t) + o(n^{-1/2}); \\ &\mathbb{P}^*(\widehat{T} \leq t) = \Phi(t) + \tfrac{\hat{K}_{3,n}}{6V_n^3\sqrt{n}}(2t^2+1)\phi(t) + o(n^{-1/2}). \end{aligned} \tag{18}$$

Proofs of (b) and (c) also show that $\hat{K}_{3,n} - \bar{\mu}_{3,n} \to 0$, yielding that $\mathbb{P}^*(\widehat{T} \leq t) = \mathbb{P}(T \leq t) + O_p(n^{-1/2})$. Then, Proposition 1 has been proven.

B. PROOF OF PROPOSITION 2

Note that $\mathbb{P}[\theta(\boldsymbol{X}) \leq \hat{\theta}(\boldsymbol{X}) - \frac{F^{-1}(1-\alpha)[\widehat{\mathbb{V}\text{ar}}(\hat{\theta}(\boldsymbol{X}))]^{\frac{1}{2}}}{\sqrt{|\mathcal{S}_n|}}] = \alpha$. as $\mathbb{P}(T \leq t) = \alpha$. So, $\hat{\theta}_{T,\boldsymbol{X}}(\alpha) = \hat{\theta}(\boldsymbol{X}) - \frac{F^{-1}(1-\alpha)[\widehat{\mathbb{V}\text{ar}}(\hat{\theta}(\boldsymbol{X}))]^{\frac{1}{2}}}{\sqrt{|\mathcal{S}_n|}}$. For the bootstrapping, $\hat{\theta}_{\widehat{T},\boldsymbol{X}}(\alpha) = \hat{\theta}(\boldsymbol{X}) - \frac{\widehat{F}^{-1}(1-\alpha)[\widehat{\mathbb{V}\text{ar}}(\hat{\theta}(\boldsymbol{X}^*))]^{\frac{1}{2}}}{\sqrt{|\mathcal{S}_n|}}$. Combing the facts that $\widehat{F}^{-1} = F^{-1} + O_p(n^{-1})$ and $\widehat{\mathbb{V}\text{ar}}(\hat{\theta}(\boldsymbol{X}^*)) = \widehat{\mathbb{V}\text{ar}}(\hat{\theta}(\boldsymbol{X})) + O_p(n^{-1})$ from [6], the proof of Proposition 2 is straightforward.