From Truth Discovery to Trustworthy Opinion Discovery: An Uncertainty-Aware Quantitative Modeling Approach

Mengting Wan[*], Xiangyu Chen[†], Lance Kaplan[‡], Jiawei Han[†], Jing Gao[§], Bo Zhao[¶]

[*] University of California, San Diego, La Jolla, CA, USA
[†] University of Illinois, Urbana-Champaign, Urbana, IL, USA
[‡] U.S. Army Research Laboratory, Adelphi, MD, USA
[§] SUNY Buffalo, Buffalo, NY, USA [¶] LinkedIn, Mountain View, CA, USA

[*]m5wan@ucsd.edu, [†]{chen338, hanj}@illinois.edu, [‡]lance.m.kaplan.civ@mail.mil, [§]jing@buffalo.edu, [¶]bo.zhao.uiuc@gmail.com

ABSTRACT

In this era of information explosion, conflicts are often encountered when information is provided by multiple sources. Traditional truth discovery task aims to identify the *truth* – the most trustworthy information, from conflicting sources in different scenarios. In this kind of tasks, truth is regarded as a fixed value or a set of fixed values. However, in a number of real-world cases, objective truth existence cannot be ensured and we can only identify single or multiple reliable facts from opinions. Different from traditional truth discovery task, we address this uncertainty and introduce the concept of *trustworthy opinion* of an entity, treat it as a *random variable*, and use its distribution to describe consistency or controversy, which is particularly difficult for data which can be numerically measured, i.e. quantitative information. In this study, we focus on the quantitative opinion, propose an uncertainty-aware approach called Kernel Density Estimation from Multiple Sources (**KDEm**) to estimate its probability distribution, and summarize trustworthy information based on this distribution. Experiments indicate that **KDEm** not only has outstanding performance on the classical numeric truth discovery task, but also shows good performance on multi-modality detection and anomaly detection in the uncertain-opinion setting.

Keywords

Truth Discovery; Source Reliability; Kernel Density Estimation

1. INTRODUCTION

In this era of information explosion, numerous claims about the same object can be collected from multiple sources. Examples include city weather information found through different websites, product rating scores collected from different customers, and gun control comments provided by different political parties. However, these claims are usually not consistent and conflicts may appear from different sources. Therefore, how to integrate and summarize conflicting claims and to find out trustworthy information from multiple sources becomes a challenge.

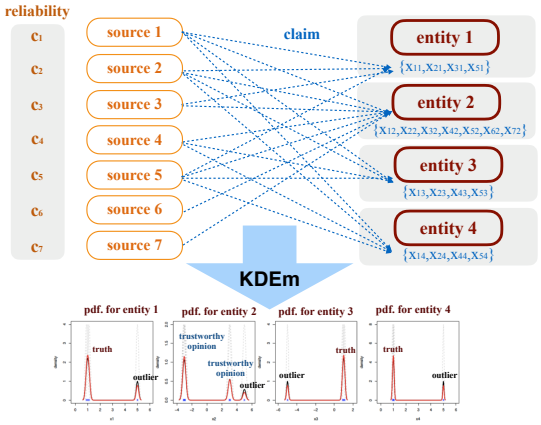


Figure 1: General Workflow: from *Truth Discovery* to *Trustworthy Opinion Discovery*.

Truth Discovery. To solve this problem, a series of truth discovery models were developed, where the concept of *truth* is implicated as a fact or a set of facts which can be *consistently agreed*. A straightforward approach to solve this problem for categorical data is to take the majority as the truth. For numeric data, mean or median can be regarded as the truth. These straightforward methods regard different sources as equally reliable, which may fail in scenarios where data are not clean enough and inputs are contaminated by unreliable sources, such as out-of-date websites, faulty devices and spam users. Therefore, several methods have been proposed to overcome this weakness by estimating source reliability and trustworthy information simultaneously [4, 5, 9, 10, 13, 15–20, 24–27].

Truth or Trustworthy Opinion? We notice that because of the objectivity of *truth*, the output for an entity from most existing truth discovery models is *a fixed value* while other pieces of information are discarded. However, the objective truth may not be found or the existence of it cannot be ensured for a number of cases. For example, the

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

DOI: http://dx.doi.org/10.1145/2939672.2939837

	Entity 1	Entity 2	Entity 3	Entity 4
Source 1	1.00	3.00	1.00	0.95
Source 2	1.10	3.10	0.90	1.00
Source 3	0.90	-3.00	-	-
Source 4	-	-3.10	1.10	1.05
Source 5	5.00	5.00	-5.00	5.00
Source 6	-	-2.90	-	-
Source 7	-	-3.05	-	-

Table 1: Example 1: A toy example for trustworthy opinion discovery.

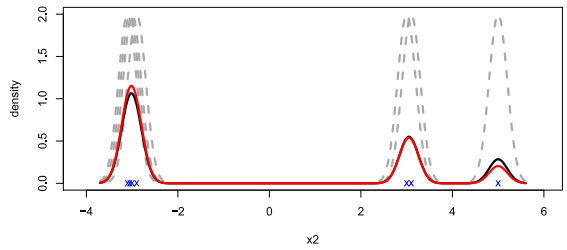


Figure 2: Probability density estimation for Entity 2 in Example 1.

exact decline time for Maya civilization remains a mystery and the number for Apple Watch sales is kept secret to the public. For such category of problems, answers of multiple versions from multiple sources stay active, which greatly invalidate the power of traditional truth discovery approaches. In these cases, we can only summarize *reliable facts from opinion claims* provided by multiple sources. Some of these entities may have only one dominant fact while others may have multiple reliable representative opinion instances. We can provide several real-world scenarios as follows.

- *The correct answer may be controversial because of the **ambiguity** of a query or **lack of certain conditions**, but dominant answers can be summarized for reference. For example, in the social sensing task, data from sensors may tend to be divided into different clusters because of unobserved conditions, and representative centers can be concluded.*
- *People's feedback regarding a product or a business may be controversial because of the **subjectivity**, but trustworthy opinion instances can be summarized. For example, in the review rating summarization task, American audience's rating distribution and Chinese audience's rating distribution may be different for an American TV show related to China, such as Marco Polo (2014), due to the cultural difference.*
- *The existence of truth cannot be found or ensured for some **open questions** and **confidential statistics**, but single or multiple promising candidate answers can be concluded. For example, the potential cause of a particular type of disease could be an open question. However, if we retrieve it from medical literature, one or more promising causes can be found.*

Since truth can only be represented by a fixed value or a set of fixed values, to model all above scenarios, we need to replace the concept of *truth* by the concept of *trustworthy opinion* (*opinion* as shorthand) of an entity. To preserve the *uncertainty* of opinion, we will regard the opinion as a *random variable* and find its distribution to describe the consistency or the controversy.

Uncertainty of Quantitative Opinion. Most truth discovery models designed for categorical data can provide a trustworthiness score to each claim and assign the one with the largest score to be the truth of this entity. This score indeed can be regarded as a reflection of probability for a claim of being chosen as truth. Therefore, it could be straightforward to extend these methods to model the distribution of categorical opinion. However, we notice that there are numerous cases where data are numeric or can be quantitatively measured. It is nontrivial to model this kind of quantitative information in an uncertainty-aware way. In existing numeric truth discovery models, it is believed that the truth is a single value and the uni-modal distribution for the truth is assumed or implicated. For entities of which trustworthy opinions are controversial, this uni-modal assumption may cause a loss of valuable information.

Example 1. *Table 1 is a toy example to illustrate uncertain quantitative opinion, which is composed of four entities. Claims are provided from seven sources. In this example, opinions of Entity 1, Entity 3 and Entity 4 may be consistent while the opinion of Entity 2 may be controversial.*

When we apply traditional numeric truth discovery models on Example 1, due to the uni-modal implication, the truth estimation for Entity 2 may shrink to a value between two modes "3" and "-3" and source reliability cannot be estimated appropriately. Even if we can specify the multi-modality of the data, it is difficult to identify the number of modes so that parametric approaches can hardly be applied to solve this problem.

To solve the problem of trustworthy *quantitative* opinion discovery from multiple sources, we need to overcome several challenges as follows. Firstly, how can we preserve the uncertainty of opinion and model the source reliability simultaneously? Secondly, if the reliable underlying opinion distribution is obtained, how can we find the truth from this if we know truth exists for an entity? Also, how can we summarize representative quantitative opinion instances if we are not sure about the truth existence?

1.1 Overview of Proposed Method

In this study, we propose an uncertainty-aware method to summarize trustworthy *quantitative* information from multiple sources. The general workflow is described as follows.

- Firstly, we estimate opinion distributions of entities which are represented as **p**robability **d**ensity **f**unctions (*pdf*). Specifically, we introduce a nonparametric model of Kernel Density Estimation from Multiple Sources (**KDEm**), which is also the core algorithm of this study.
- Secondly, if truth exists, we estimate the truth from obtained opinion distribution; if truth existence cannot be ensured, then we report representative opinion instance(s) based on estimated opinion distribution.

Figure 1 describes the workflow which is applied on aforementioned Example 1.

The philosophy of **KDEm** is similar to the standard Kernel Density Estimation (**KDE**) [14]. To model the shape of a probability density function (*pdf*), a straightforward approach is drawing histogram, which is usually not smooth enough for numeric data. However, by applying kernel technique, we can add continuity over bins and obtain a smooth *pdf* estimation – kernel density estimation (**KDE**). Using kernel, we can transform each claim from a real value to a single component function. For each entity, standard **KDE** is to find a function which is similar to all the component functions. Then the multi-modality of opinion distribution can be preserved through this technique. Below is an example illustrating this idea.

Example 2. *Each single component for Entity 2 in Table 1 is plotted in Figure 2 as the grey dotted line. The standard **KDE** for the opinion of each entity in Table 1 is plotted in Figure 2 as the black solid line, and the controversial opinion can be preserved in this way (represented as peaks).*

	Truth Discovery	Trustworthy Opinion Discovery
input	entities; claims; sources.	
target	truth (fixed value)	trustworthy opinion (random variable)
output	*value* for truth	*probability distribution* for opinion - if truth exists: value for truth - otherwise: single or multiple representative values
source reliability?	Yes	
multi-modality detection?	No	Yes
Anomaly detection?	No	Yes
Robust to outliers? (numeric data)	No	Yes

Table 2: Truth Discovery v.s. Trustworthy Opinion Discovery.

We also notice that Source 5 consistently contaminates the data but in **KDE** model, component functions are equally weighted. Thus in this study, we believe that reliable source provides trustworthy claims. Then our **KDEm** can be regarded as an optimization framework to find a target function which can minimize the weighted difference between the target function and each single component function, where the weight reflects corresponding source reliability. Here we illustrate how the proposed **KDEm** is able to capture source reliability.

Example 3. *In Figure 2, the output from **KDEm** is plotted as the red solid line. We notice that compared with the standard **KDE**, **KDEm** can reduce the effect of untrustworthy data provided by Source 5.*

Once the reliable *pdf* of the quantitative opinion is obtained from **KDEm**, we can cluster claims based on this function. Within each cluster, we can regard the mode or the claim with the largest *pdf* value as a representative candidate. If we know truth exists for the entity, the most trustworthy candidate will be reported as the truth and others will be treated as outliers. Through this approach, **KDEm** is robust to outliers and can naturally detect outliers. Thus different from other numeric truth discovery models, no additional outlier detection procedure is needed for **KDEm**. If the truth existence cannot be ensured, by setting a confidence threshold, we can report single or multiple representative values as trustworthy opinion instances, identify uni-modality/multi-modality and detect anomaly observations.

Contributions of this Study. The classical truth discovery task and our trustworthy opinion discovery task are compared in Table 2. Generally, these two kinds of problems take the same input and both involve source reliability. However, their output formats are different and ideally an trustworthy opinion discovery model can be compatible with classical truth discovery task when truth existence can be ensured. Now we conclude contributions of this study as follows.

- Different from previous truth discovery models, we raise a new but closely related problem – *trustworthy opinion discovery*. We replace the concept of *truth* by the concept of *trustworthy opinion*, model the uncertainty of quantitative opinion and regard the opinion as a random variable;
- A nonparametric approach **KDEm** is proposed to estimate opinion distribution and source reliability score simultaneously, which can model different shapes of density functions and perceive multiple modes;
- **KDEm** is compatible with traditional numeric truth discovery task, and could be significantly robust to outliers;
- Based on the opinion distribution estimation from **KDEm**, we can summarize one or more representative values, distinguish controversial entities from consistent entities (unimodal/multi-modal detection) and identify abnormal claims (anomaly detection).

The rest of the paper is organized as follows. We illustrate definitions and problem formulation in Section 2. Section 3 describes our model for this trustworthy quantitative opinion discovery task. Section 4 presents our experimental results. We then introduce related works in Section 5 and provide conclusions and future directions in Section 6.

2. BACKGROUND

In this section, we formally define the trustworthy opinion discovery task. We first define some basic terms:

Definition 2.1.

- *An **entity** is an object of interest.*
- *A **claim** is a value provided by a **source** for an entity.*
- *A **trustworthy opinion** is a random variable whose distribution describes the trustworthy information of an entity.*
- *A **truth** is a fixed value regarding an entity which can be consistently agreed. If **truth** exists for an entity, it can be distinguished based on the distribution of trustworthy opinion.*
- *The **representative value**(s) of an opinion could be one or more significant trustworthy values summarized based on the opinion distribution.*
- *The **confidence of a representative value** is a score that measures the significance level of the representative value of an opinion. Higher confidence indicates this representative value is more trustworthy and vice versa.*
- *A **source reliability score** describes the possibility of a source providing trustworthy claims. Higher source reliability score indicates that the source is more reliable and vice versa.*

Notice that in this study, we only discuss the quantitative opinion with single value setting, which means the trustworthy opinion is a numeric random variable. Then we define the trustworthy opinion discovery task as follows:

Definition 2.2. ***(Trustworthy Opinion Discovery)*** *For a set of entities $\mathcal{N}$ of interest, claims are collected from a set of sources $\mathcal{S}$. The uncertainty-aware trustworthy opinion discovery task is to estimate the probability density function of the trustworthy opinion of each entity, and identify the reliability level of each source simultaneously.*

To better understand the estimated opinion distribution, we can summarize the representative values and associated confidence scores based on the estimated probability density function of the opinion. Details about this procedure will be introduced in Section 3.2.

All the notations used in this study has been summarized in Table 3.

3. METHOD

Generally, the method for uncertainty-aware quantitative trustworthy information summarization can be divided into two steps: 1) estimating the density function of the opinion of each entity; and 2) summarizing the trustworthy information based on estimated opinion distribution.

Notation	Definition
N_i	the i-th entity
$\mathcal{N}$	$\mathcal{N} := \{N_1, ..., N_n\}$; a set of n entities
S_j	the j-th source
$\mathcal{S}$	$\mathcal{S} := \{S_1, ..., S_m\}$; a set of m sources
c_j	the reliability score of the j-th source
$\mathcal{N}_j$	the set of index of entities where claims are provided by the j-th source
n_j	the number of entities where claims are provided by the j-th source
$\mathcal{S}_i$	the set of index of sources who provide claims for the i-th entity
m_i	the number of sources who provides claims for the i-th entity
$\boldsymbol{x}_{ij}$	the claim provided by the j-th source for the i-th entity
$\mathcal{X}_i$	$\mathcal{X}_i = \{\boldsymbol{x}_{ij}\}_{j\in\mathcal{S}_i}$; the set of m_i claims for the i-th entity
$\mathcal{X}$	$\mathcal{X} = \cup_{i=1}^{n}\mathcal{X}_i$; the set of claims for all the entities
$\boldsymbol{t}_i$	the trustworthy opinion for the i-th entity, which is a random variable
f_i	the probability density function of $\boldsymbol{t}_i$
$\boldsymbol{t}^*_{ik}$	the k-th representative value of $\boldsymbol{t}_i$
k_i	the number of representative values of $\boldsymbol{t}_i$
$\mathcal{T}^*_i$	$\mathcal{T}^*_i := \{\boldsymbol{t}^*_{i1}, ..., \boldsymbol{t}^*_{ik_i}\}$; the set of k_i representative values of the trustworthy opinion $\boldsymbol{t}_i$

Table 3: Notation

3.1 Kernel Density Estimation from Multiple Sources

In this section, we first introduce the intuition and a density estimation method without distinguishing sources. Then we introduce our model and the algorithm.

3.1.1 Intuition: from a Real Coordinate Space to a Function Space

Suppose the claim set for the i-th entity is denoted by $\{\boldsymbol{x}_{ij} \in \mathbb{R}^d, j \in \mathcal{S}_i\}$. For the traditional truth discovery task, a straightforward estimation of the truth is the sample mean. By introducing the concept of source reliability, the format of weighted sample mean is applied in several existing numeric truth discovery methods [9, 10]. Here the weights correspond to source reliability scores.

As discussed before, in our uncertain-opinion setting, to model the uncertainty of opinion, we need to map truths and claims from real values/vectors to functions. Therefore, we define this mapping for the i-th entity as

$$\begin{aligned} \Phi_i : \mathbb{R}^d &\to \mathcal{H}_i \\ \boldsymbol{x} &\mapsto K_{h_i}(\cdot, \boldsymbol{x}) := \Phi_i(\boldsymbol{x}), \end{aligned} \tag{1}$$

where K_{h_i} is a translation invariant, symmetric, positive semi-definite kernel function with bandwidth h_i ($h_i > 0$) for the i-th entity. K_{h_i} needs to satisfy $K_{h_i}(\cdot, \boldsymbol{x}) \geq 0$ and $\int K_{h_i}(\boldsymbol{x}', \boldsymbol{x})d\boldsymbol{x}' = 1$, so that it can be ensured as a probability density. A typical kernel example is **Gaussian kernel**:

$$K_{h_i}(\boldsymbol{x}', \boldsymbol{x}) = (\frac{1}{\sqrt{2\pi}h_i})^d \exp(-(\frac{\|\boldsymbol{x}' - \boldsymbol{x}\|}{h_i})^2), \tag{2}$$

which is used in all the experiments in this study. If Gaussian kernel is applied, we notice that the function transformation of $\boldsymbol{x}_{ij}$, $\Phi_i(\boldsymbol{x}_{ij}) = K_{h_i}(\cdot, \boldsymbol{x}_{ij})$, is a density function of Gaussian distribution.

By applying this kind of mapping, we have following analogies of previous sample mean and weighted sample mean:

$$\overbrace{\frac{1}{m_i}\sum_{j\in\mathcal{S}_i}\boldsymbol{x}_{ij}}^{\text{sample mean}} \mapsto \overbrace{\frac{1}{m_i}\sum_{j\in\mathcal{S}_i}\Phi_i(\boldsymbol{x}_{ij})}^{\text{sample mean function, i.e. KDE}} \tag{3}$$

$$\overbrace{\frac{1}{m_i}\sum_{j\in\mathcal{S}_i} w_{ij}\boldsymbol{x}_{ij}}^{\text{weighted sample mean}} \mapsto \overbrace{\frac{1}{m_i}\sum_{j\in\mathcal{S}_i} w_{ij}\Phi_i(\boldsymbol{x}_{ij})}^{\text{weighted sample mean function}} \tag{4}$$

where $\Phi_i(\boldsymbol{x}_{ij}) = K_{h_i}(\cdot, \boldsymbol{x}_{ij})$ and $\sum_{j\in\mathcal{S}_i} w_{ij} = 1$. The sample mean function in (3) can be written as

$$\hat{f}_i(\boldsymbol{t}_i) = \frac{1}{m_i}\sum_{j\in\mathcal{S}_i} K_{h_i}(\boldsymbol{t}_i, \boldsymbol{x}_{ij}), \tag{5}$$

which is the standard Kernel Density Estimation (**KDE**) [14] of the opinion $\boldsymbol{t}_i$. By considering source trustworthiness, we have the extended weighted sample mean function in (4). The major task in our **KDEm** model is to find the specific *pdf* estimation in this format.

In preparation for subsequent analysis, we need to look at the kernel technique in detail and define inner product, norm and distance for this function space. Each positive semi-definite kernel K_{h_i} is associated with a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_i$ [1]. For $\boldsymbol{x} \in \mathbb{R}^d$, we have $\Phi_i(\boldsymbol{x}) = K_{h_i}(\cdot, \boldsymbol{x}) \in \mathcal{H}_i$, $\frac{1}{m_i}\sum_{j\in\mathcal{S}_i}\Phi_i(\boldsymbol{x}_{ij}) \in \mathcal{H}_i$ and $\frac{1}{m_i}\sum_{j\in\mathcal{S}_i} w_{ij}\Phi_i(\boldsymbol{x}_{ij}) \in \mathcal{H}_i$.

Inner Product. Based on the reproducing property, for $g \in \mathcal{H}_i$, $\boldsymbol{x} \in \mathbb{R}^d$, we have the definition of inner product [1]

$$\langle \Phi_i(\boldsymbol{x}), g\rangle_{\mathcal{H}_i} = g(\boldsymbol{x}). \tag{6}$$

Specially, by taking $g = K_{h_i}(\cdot, \boldsymbol{x}') = \Phi_i(\boldsymbol{x}')$, we have

$$\left\langle \Phi_i(\boldsymbol{x}), \Phi_i(\boldsymbol{x}')\right\rangle_{\mathcal{H}_i} = K_{h_i}(\boldsymbol{x}, \boldsymbol{x}'). \tag{7}$$

Norm and Distance. Then we have the definition of the norm $\|\cdot\|$:

$$\|f\|_{\mathcal{H}_i} = \sqrt{\langle f, f\rangle_{\mathcal{H}_i}}, \tag{8}$$

and the definition of distance between two functions $f, g \in \mathcal{H}_i$:

$$\|f - g\| = \sqrt{\|f\|^2_{\mathcal{H}_i} - 2\langle f, g\rangle + \|g\|^2_{\mathcal{H}_i}}. \tag{9}$$

3.1.2 Kernel Density Estimation from Multiple Sources (KDEm)

We now define our model – Kernel Density Estimation from Multiple Sources (**KDEm**), by introducing the source weight and minimizing the loss on different entities together.

Particularly, we need to find a set of functions $f_i \in \mathcal{H}_i$, $i = 1, ..., n$ and a set of numbers $c_j \in \mathbb{R}^+$, $j = 1, ..., m$, which can minimize the total loss function

$$J(f_1, ..., f_n; c_1, ..., c_m) = \sum_{i=1}^{n}\frac{1}{m_i}\sum_{j\in\mathcal{S}_i} c_j\|\Phi_i(\boldsymbol{x}_{ij}) - f_i\|^2_{\mathcal{H}_i} \tag{10}$$

where m_i is the number of provided claims for the i-th entity, and $c_1, ..., c_m$ satisfy

$$\sum_{j=1}^{m} n_j \exp(-c_j) = 1. \tag{11}$$

where n_j is the number of claims provided by S_j. Suppose $\hat{f}_i^{kdem}$ is the output for f_i from this framework. Then $\hat{f}_i^{kdem}$ is defined as the density estimation for $\boldsymbol{t}_i$, the trustworthy opinion of the i-th entity $(i = 1, ..., n)$.

In (10), c_j reflects the trustworthiness level of source S_j and $\|\Phi_i(\boldsymbol{x}_{ij}) - f_i\|_{\mathcal{H}_i}$ measures the distance between the opinion density f_i and $\Phi_i(\boldsymbol{x}_{ij})$, the function transformation of the claim $\boldsymbol{x}_{ij}$ provided by S_j. If S_j is reliable, it will give large penalty to the distance and vice versa. We use the constraint (11) to ensure the number of solutions for $c_1, ..., c_m$ is finite and this optimization problem is convex if $f_1, ..., f_n$ are given. In (11), n_j is used to model the involvement level of source S_j.

To minimize the total loss function (10) with constraint (11), we further convert the problem into an optimization problem without constraint. That is to find a set of functions $f_i \in \mathcal{H}_i$ for $i = 1, ..., n$, a set of numbers $c_j \in \mathbb{R}^+$ for $j = 1, ..., m$, and a real number λ to minimize the new loss function

$$\begin{aligned} &Q(f_1, ..., f_n; c_1, ..., c_m; \lambda) \\ =&J(f_1, ..., f_n; c_1, ..., c_m) + \lambda(\sum_{j=1}^{m} n_j \exp(-c_j) - 1). \end{aligned} \tag{12}$$

For $\mathbb{R}^d$ and the function $F : \mathbb{R}^d \to \mathbb{R}$, the Gateaux differentials of F at $\boldsymbol{x} \in \mathbb{R}^d$ with incremental $\boldsymbol{h} \in \mathbb{R}^d$ is

$$dF(\boldsymbol{x}; \boldsymbol{h}) = \lim_{\alpha \to 0} \frac{F(\boldsymbol{x} + \alpha \boldsymbol{h})}{\alpha} = \frac{d}{d\alpha} F(\boldsymbol{x} + \alpha \boldsymbol{h})\Big|_{\alpha=0}, \tag{13}$$

if the limit exists for all $\boldsymbol{h} \in \mathbb{R}^d$. Then a necessary condition for F to achieve a minimum at $\boldsymbol{x}_0$ is $dF(\boldsymbol{x}_0; \boldsymbol{h}) = 0$ for $\forall \boldsymbol{h} \in \mathbb{R}^d$. We thus have the following lemma:

Lemma 3.0.1. *For $\forall i \in \{1, ..., n\}$, given $\{c_1, ..., c_m \in \mathbb{R}^+\}$, $\{f_j \in \mathcal{H}_j | j = 1, ..., n, j \neq i\}$ and $\lambda \in \mathbb{R}$, the Gateaux differential of Q at $f_i \in \mathcal{H}_i$ with incremental $\boldsymbol{h} \in \mathcal{H}_i$ can be given by*

$$\begin{aligned} &d_i Q(f_1, ..., f_n; c_1, ..., c_m; \lambda) \\ =&\frac{d}{d\alpha} Q(f_1, ..., f_i(\boldsymbol{x} + \alpha \boldsymbol{h}), ..., f_n; c_1, ..., c_m; \lambda)\Big|_{\alpha=0} \\ =&- \langle V_i(f_i), \boldsymbol{h} \rangle_{\mathcal{H}_i}, \end{aligned} \tag{14}$$

where $V_i(f_i) = \frac{2}{m_i} \sum_{j \in \mathcal{S}_i} c_j (\Phi_i(\boldsymbol{x}_{ij}) - f_i)$.

We can prove this lemma by applying similar technique in [7]. Given $c_1, ..., c_m \in \mathbb{R}^+$, a necessary condition for $f_i = \hat{f}_i^{kdem}$ is $V_i(f_i) = \frac{2}{m_i} \sum_{j \in \mathcal{S}_i} c_j (\Phi_i(\boldsymbol{x}_{ij}) - f_i) = \boldsymbol{0}$. By solving it, we have the following theorem for $f_i \in \mathcal{H}_i$:

Theorem 3.1. *Suppose $c_1, ..., c_m \in \mathbb{R}^+$ are fixed, the estimation for $f_i \in \mathcal{H}_i$, $i = 1, ..., n$ can be given by a weighted kernel density estimation*

$$\hat{f}_i^{kdem} = \sum_{j \in \mathcal{S}_i} w_{ij} \Phi_i(\boldsymbol{x}_{ij}), \tag{15}$$

where $w_{ij} = c_j / (\sum_{j' \in \mathcal{S}_i} c_{j'})$.

Notice that if sources are equally reliable, we have $c_1 = ... = c_m$ and the estimated *pdf* from (15) is the same output from standard **KDE**.

If $f_i \in \mathcal{H}_i, \forall i = 1, .., n$ are fixed, by solving the equations $\frac{\partial}{\partial c_j} Q = 0$ and $\frac{\partial}{\partial \lambda} Q = 0$, and calculating $\frac{\partial^2}{\partial^2 c_j} Q$ for $j = 1, ..., m$, we have the following theorem for $c_1, ..., c_m \in \mathbb{R}^+$:

Theorem 3.2. *Suppose $f_i \in \mathcal{H}_i, i = 1, ..., n$ are fixed, the objective problem Q is a convex optimization problem. The optimal solution for $c_j \in \mathbb{R}^+$, $j = 1, ..., m$ is*

$$c_j = -\log \left(\frac{\frac{1}{n_j} \sum_{i \in \mathcal{N}_j} \frac{1}{m_i} \|\Phi_i(\boldsymbol{x}_{ij}) - f_i\|_{\mathcal{H}_i}^2}{\sum_{j'=1}^{m} \sum_{i \in \mathcal{N}_{j'}} \frac{1}{m_i} \|\Phi_i(\boldsymbol{x}_{ij'}) - f_i\|_{\mathcal{H}_i}^2} \right). \tag{16}$$

Therefore, we can apply a block coordinate descent [2] iterative method, which can keep reducing the total loss function (10), to obtain the estimated densities $\hat{f}_i, i = 1, ..., n$ and source weight scores $c_j, j = 1, ..., m$. This method is concluded as Algorithm 1.

Algorithm 1 KDEm Algorithm

(a) Initialize $c_1^{(0)} = ... = c_j^{(0)} = ... = c_m^{(0)}$;

(b) Update $\hat{f}_i$ by $\hat{f}_i^{(k+1)} = \sum_{j \in \mathcal{S}_i} w_{ij}^{(k)} \Phi_i(\boldsymbol{x}_{ij})$,
where $w_{ij}^{(k)} = \frac{c_j^{(k)}}{\sum_{j' \in \mathcal{S}_i} c_{j'}^{(k)}}$, $i = 1, ..., n$;

(c) Update c_j by
$$c_j^{(k+1)} = -\log \left(\frac{\frac{1}{n_j} \sum_{i \in \mathcal{N}_j} \frac{1}{m_i} \|\Phi_i(\boldsymbol{x}_{ij}) - \hat{f}_i^{(k+1)}\|_{\mathcal{H}_i}^2}{\sum_{j'=1}^{m} \sum_{i \in \mathcal{N}_{j'}} \frac{1}{m_i} \|\Phi_i(\boldsymbol{x}_{ij'}) - \hat{f}_i^{(k+1)}\|_{\mathcal{H}_i}^2} \right);$$
$j = 1, ..., m$

(d) Repeat (b) and (c) until the total loss $J(f_1, ..., f_n; c_1, ..., c_m)$ showed in (10) does not change.

The general principle of Algorithm 1 is that we start with the opinion density functions obtained from standard **KDE** and then iteratively update the opinion distributions and the source reliability scores. If obtained opinion densities are closer to the real trustworthy opinion distributions, then preciser source reliability scores can be obtained based on (16). On the other hand, if the updated source reliability scores are more accurate, then we can obtain preciser trustworthy opinion densities based on (15). Therefore, these two updating procedures can mutually enhance each other.

Specifically in Algorithm 1, since

$$\hat{f}_i^{(k+1)} = \sum_{j \in \mathcal{S}_i} w_{ij}^{(k)} \Phi_{ij}$$

where Φ_{ij} is the shorthand of $\Phi_i(\boldsymbol{x}_{ij})$, we have

$$\begin{aligned} &\|\Phi_i(\boldsymbol{x}_{ij}) - \hat{f}_i^{(k+1)}\|_{\mathcal{H}_i}^2 = \|\Phi_{ij} - \sum_{j' \in \mathcal{S}_i} w_{ij'}^{(k)} \Phi_{ij'}\|_{\mathcal{H}_i}^2 \\ =&\|\Phi_{ij}\|_{\mathcal{H}_i}^2 - 2 \sum_{l \in \mathcal{S}_i} w_{il}^{(k)} \langle \Phi_{ij}, \Phi_{il} \rangle + \sum_{l,l' \in \mathcal{S}_i} w_{il}^{(k)} w_{il'}^{(k)} \langle \Phi_{il}, \Phi_{il'} \rangle \\ =&K_{h_i}(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij}) - 2 \sum_{l \in \mathcal{S}_i} w_{il}^{(k)} K_{h_i}(\boldsymbol{x}_{ij}, \boldsymbol{x}_{il}) + \sum_{l,l' \in \mathcal{S}_i} w_{il}^{(k)} w_{il'}^{(k)} K_{h_i}(\boldsymbol{x}_{il}, \boldsymbol{x}_{il'}) \end{aligned}$$

In our model, h_i can be decided either based the data or based on prior knowledge. Details about bandwidth selection will be introduced in the experiment part. Here we show how the algorithm works on the aforementioned example.

Example 4. *For data in Table 1, we start with the equally weighted sources and calculate the source reliability score (c_j) in each iteration in Algorithm 1. The results are showed in Figure 3, from which we notice that the source reliability score of Source 5 can be constantly reduced while others can be constantly increased until convergence in **KDEm**. Here h_i is set to be:*

$$MAD_i = median\{\|\boldsymbol{x}_{ij} - median\{\boldsymbol{x}_{ij'}\}_{j' \in \mathcal{S}_i}\|\}_{j \in \mathcal{S}_i}.$$

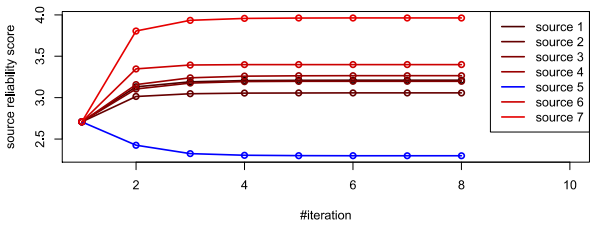


Figure 3: Source reliability score c_j in each iteration for Example 1.

3.1.3 *Time Complexity and Practical Issues*

In each iteration, for each entity, the most time consuming part is to compute $\|\Phi_i(\boldsymbol{x}_{ij}) - \hat{f}_i^{(k+1)}\|^2_{\mathcal{H}_i}$, which takes $O(m_i^2)$ time, where m_i is the number of claims for the i-th entity. Thus it takes $O(\sum_{i=1}^n m_i^2)$ time for each iteration and $O(k\sum_{i=1}^n m_i^2)$ for the whole **KDEm** model, where k is the number of iterations ($k < 10$ in our experiments).

In some real cases, although data are numeric and the number of claims is significantly large, the possible values are limited. For example, the values of rating scores are usually integers from 1–5 or from 1–10. In such cases, for each entity, we can easily map these claims to corresponding values. Then we can compute the kernel basis $K(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'})$ and the distance $\|\Phi_i(\boldsymbol{x}_{ij}) - \hat{f}_i^{(k+1)}\|^2_{\mathcal{H}_i}$ only for mapped values. The time cost for each iteration thus becomes $\sum_{i=1}^n v_i^2$, where v_i is the number of claimed values for the i-th entity.

3.2 Trustworthy Information Summarization Based on the Opinion Distribution

Once the opinion density estimation $\hat{f}_i^{kdem}$ for each entity N_i is obtained, we can use DENCLUE 2.0 [6] to cluster claims $\{\boldsymbol{x}_{ij}, j \in \mathcal{S}_i\}$ and calculate the center of each cluster based on the opinion distribution. Then from these clusters, we can summarize the representative values and corresponding confidence values based on different user preferences.

Clustering Claims. DENCLUE 2.0 [6], a hill-climbing procedure which assigns each claim to its nearest mode based on the density function, is applied in this part. Specifically, taking Gaussian kernel as example, the gradient of $\hat{f}_i^{kdem}(\boldsymbol{t}_i)$ is given by

$$\nabla \hat{f}_i^{kdem}(\boldsymbol{t}_i) = \frac{1}{h_i^{d+1}} \sum_{j \in \mathcal{S}_i} w_{ij} K_{h_i}(\boldsymbol{x}_{ij}, \boldsymbol{t}_i) \cdot (\boldsymbol{x}_{ij} - \boldsymbol{t}_i). \quad (17)$$

By setting it to zero, we obtain an update rule:

$$\boldsymbol{t}_i^{(l+1)} = \frac{\sum_{j \in \mathcal{S}_i} w_{ij} K_{h_i}(\boldsymbol{x}_{ij}, \boldsymbol{t}_i^{(l)}) \boldsymbol{x}_{ij}}{\sum_{j \in \mathcal{S}_i} w_{ij} K_{h_i}(\boldsymbol{x}_{ij}, \boldsymbol{t}_i^{(l)})} \quad (18)$$

For each entity N_i, the procedure starts at each claim $\boldsymbol{x}_{ij}$ and iteratively update it based on (18) until convergence. For claims which converge to the same mode $\hat{\boldsymbol{x}}_{ik}$, we cluster them together. The cluster is denoted as $\mathcal{C}_{ik}$ and the confidence of this cluster is defined as $c_{ik} = \sum_{j:\boldsymbol{x}_{ij} \in \mathcal{C}_{ik}} w_{ij}$.

Summarizing Trustworthy Information. We first summarize the representative candidate value within each cluster. Then we screen these candidates based on certain criteria and report representative values of the opinion based on different user preferences. Here we introduce two sets of user preferences regarding these two steps respectively as follows.

- **"Discrete" vs. "Continuous".** Although our model is designed for numeric data, in real cases, e.g., "the number of Solar System planets", numeric claims may share discrete property as well and users may believe that a representative value should be from provided claims. In this case, for each entity N_i, within each cluster $\mathcal{C}_{ik}$, the claim with the largest density value is regarded as a representative candidate (**"Discrete"**):

$$\hat{\boldsymbol{t}}_{ik}^* = \arg \max_{\boldsymbol{x}_{ij} \in \mathcal{C}_{ik}} \hat{f}_i^{kdem}(\boldsymbol{x}_{ij}). \quad (19)$$

  However, if users believe that a representative candidate may not be claimed or observed by any sources, then the associated mode $\hat{\boldsymbol{x}}_{ik}$ can be regarded as a representative candidate (**"Continuous"**):

$$\hat{\boldsymbol{t}}_{ik}^* = \hat{\boldsymbol{x}}_{ik}. \quad (20)$$

- **"Single" vs. "Multiple".** If users believe truth exists for an entity, we only report the candidate with largest associated cluster confidence as the truth. Thus the set of reported single truth is (**"Single"**)

$$\mathcal{T}_i^* = \{\arg \max_{\hat{\boldsymbol{t}}_{ik}^*} c_{ik}\}. \quad (21)$$

  However, as we discussed before, if the truth existence cannot be ensured, then single or multiple representative values of opinion may be reported. If we are given a threshold $thr \geq 0$, then we only keep those candidates whose confidences are larger than thr and re-normalize their confidence scores. In this case, the set of reported representative values of opinion is (**"Multiple"**)

$$\mathcal{T}_i^* = \{\hat{\boldsymbol{t}}_{ik}^* | c_{ik} > thr\}. \quad (22)$$

  In the above two scenarios, we mark those claims within deleted candidates' associated clusters as outliers or anomaly observations.

4. EXPERIMENTS

In this section, we test our proposed model **KDEm** on several synthetic datasets and real world applications[1] These experiments can be categorized as two kinds of tasks:

1. Traditional truth discovery from contaminated data (single truth existence can be ensured)
2. Multi-modality detection and anomaly detection (truth existence cannot be ensured).

4.1 Traditional Numeric Truth Discovery

As discussed before, **KDEm** is compatible with traditional single truth discovery task and can be more robust to outliers compared with traditional methods. Therefore, in this section, we conduct several sets of experiments to verify the capability and superiority of **KDEm** regarding this task.

Datasets. A set of synthetic datasets `Synthetic(unimodal)` and a real world dataset `Population(outlier)` [16] are used for this task.

- `Synthetic(unimodal)` is a set of one-dimensional ($d = 1$) synthetic datasets and is generated as follows. For each single dataset in `Synthetic(unimodal)`, we generate 100 entities, 200 candidate sources and the reflection of their associated reliability scores σ_j^2, $j = 1, .., 200$. We mark $200 \times p$ sources as

[1]Data and code for this paper can be accessed through: https://github.com/MengtingWan/KDEm.

Dataset	#entity	#source	#claim	time cost
Population(outlier)	1124	2344	4008	0.2740s
Tripadvisor:				
(overall)	1759	145,291	175,766	25.85s
(value)	1759	121,480	144,128	18.88s
(rooms)	1759	122,990	146,234	19.54s
(location)	1759	107,182	124,145	15.10s
(cleanliness)	1759	122,995	146,213	18.86s
(check in)	1759	107,271	124,259	16.99s
(service)	1759	120,801	142,991	20.25s
(business service)	1759	74,227	83,670	9.356s

(Here time cost is the average time for each iteration in **KDEm** and based on seconds.)

Table 4: Basic statistics of Population(outlier) and Tripadvisor datasets and time cost from KDEm on these datasets.

"unreliable" and the remaining $200 \times (1-p)$ sources as "reliable". If a source S_j is reliable, we generate $\sigma_j \sim U(0.01, 0, 05)$; if S_j is unreliable, σ_j is generated from $U(1, 5)$. For each entity N_i, we generate the number of claims m_i from Possion distribution $\mathcal{P}(\lambda)$ and randomly select m_i sources to provide claims for this entity. We only set one ground-truthed opinion value $t_i^* = 1$ for each entity N_i and the selected source S_j provides a claim x_{ij} from Gaussian distribution $N(t_i^*, \sigma_j^2)$.

By doing so, we notice that unreliable sources may be significantly unreliable and their claims are likely to be extreme values. The parameter p here indicates the portion of unreliable sources and λ indicates the average number of claims for each entity. We test our model for $p = 0.2$ and $\lambda = 3, 5, 7, 9$. To reduce the random error, we generate 50 datasets for each pair of parameters and report the average MAE and RMSE.

- The Population(outlier) dataset is about the Wikipedia edit history regarding city population in given years. This dataset is originally published by the author of [16] and has been studied in some truth discovery studies [9, 24]. We remove some obviously-wrong claims which are more than 10^8, keep only the latest claim for the same source and the same entity, and remove entities whose claims are all the same. However, different from previous studies, we didn't apply any additional outlier detection procedures and treat the original contaminated dataset as input. The input dataset contains 4008 claims for 1124 entities from 2344 sources. Among these entities, 259 are randomly selected to be labeled with true populations. Basic statistics of this dataset are shown in Table 4. Each entity may contain outliers but the truth existence can be ensured.

For both of them, we normalize the original claims $\{\boldsymbol{x}_{ij}\}_{j\in\mathcal{S}_i}$ by its mean ($\bar{\boldsymbol{x}}_i = \sum \boldsymbol{x}_{ij}/m_i$) and standard deviation ($sd_i = \sqrt{\sum \|\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i\|^2/m_i}$). Then we use the normalized z-score ($\{(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)/sd_i\}$ as input for our model and all the baseline methods. When we obtain the output, we use the denormalized truths ($sd_i \times \boldsymbol{t}_i + \bar{\boldsymbol{x}}_i$) for evaluation.

Performance Measures. For this task, we assume that truth existence can be ensured. Thus for each entity we have only one real truth $\boldsymbol{t}_i^*$ and one estimated value $\hat{\boldsymbol{t}}_i^*$. User preference for this kind of experiments should be "Discrete"+"Single" or "Continuous"+"Single" and we try both in our experiments. We can use the Mean Absolute Error (MAE) and Rooted Mean Squared Error (RMSE) to measure the performance of models, which are defined as

- $MAE = \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{t}_i^* - \hat{\boldsymbol{t}}_i^*\|$;
- $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{t}_i^* - \hat{\boldsymbol{t}}_i^*\|^2}$.

Smaller MAE or RMSE indicates better performance.

Baselines. In addition to our model **KDEm**, we conduct standard kernel density estimation (**KDE**) [14] and robust kernel density estimation (**RKDE**) with Hampel's loss function [7] as two baselines. **RKDE** is a state-of-art M-estimation based kernel density estimation method which is more robust with outliers than standard **KDE**. For **KDE**, **RKDE** and **KDEm**, Gaussian kernel is applied and h_i is set to be the Median Absolute Deviation (MAD) of data $\mathcal{X}_i = \{\boldsymbol{x}_{ij}\}_{j\in\mathcal{S}_i}$:

$$MAD_i = median\{\|\boldsymbol{x}_{ij} - median\{\boldsymbol{x}_{ij'}\}_{j'\in\mathcal{S}_i}\|\}_{j\in\mathcal{S}_i}. \quad (23)$$

In practice, we will add a smooth item $10^{-10}sd_i$ to this value and h_i can be given by the modified MAD, which indicates

$$h_i = MAD_i^* = MAD_i + 10^{-10}sd_i. \quad (24)$$

Then we have $MAD_i^* = 0$ iff $\{\boldsymbol{x}_{ij}\}_{j\in\mathcal{S}_i}$ are all the same. For these three models, we conduct experiments based on both "Discrete"+"Single" preference (**KDEm_d**, **KDE_d**, **RKDE_d**) and "Continuous"+"Single" preference (**KDEm_c**, **KDE_c**, **RKDE_d**). Since truth existence can be ensured, we could apply several state-of-art truth discovery models on these datasets. Particularly, the following models are applied as additional baselines on both Synthetic(unimodal) and Population(outlier): **Mean**, **Median**, **TruthFinder** [23], **AccuSim** [3], **GTM** [24], **CRH** [10] and **CATD** [9]. Details about these methods will be introduced in Section 5. For Population(outlier), in addtion to these numeric truth discovery models, we add another baseline **Voting** where data are regarded as categorical and the majority value is assigned as the truth.

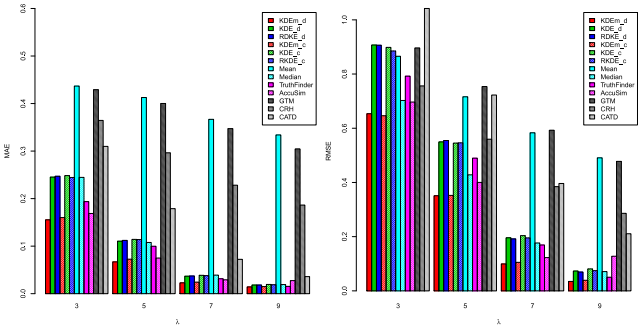


Figure 4: Results of experiments on synthetic unimodal datasets Synthetic(uni).

Results. Results of experiments on Synthetic(unimodal) are showed in Figure 4. From Figure 4, we can conclude that **KDEm_d** and **KDEm_c** generally outperform other baselines based on MAE and RMSE. We notice that **KDEm_d** and **KDEm_c** always have better performance than **KDE_d**, **RKDE_d** and **KDE_c**, **RKDE_c**, which indicate that source quality is important in our uncertain-opinion assumption. We notice that results from traditional numeric truth discovery models **GTM**, **CRH**, and **CATD** are not very good while results from **TruthFinder** and **AccuSim** are better, which are originally designed for categorical data and extended to handle numeric claims. One possible reason could be for **TruthFinder** and **AccuSim**, claims are regarded as separated facts so that the effect of outlier can be alleviated if no more trustworthy claim supports it. However, in other numeric truth discovery models, truth is regarded as a fixed value. Since real-value based distance is usually sensitive to extreme values, additional outlier detection is always needed for those methods. However, if truth is regarded as a random variable and its density function is estimated by kernel

Method	MAE	RMSE
KDEm_d	**1547**	**8884**
KDE_d	1630	8900
RKDE_d	1687	9093
KDEm_c	1875	9912
KDE_c	2024	10408
RKDE_c	2096	10643
Mean	200917	1136605
Median	11075	129850
Voting	18813	259066
TruthFinder	1551	8892
AccuSim	20819	259948
GTM	317444	1989964
CRH	219596	1289422
CATD	53750	304781

Table 5: Results of experiments on the `Population(outlier)` dataset.

methods, the effect of those extreme values can be weaken since we are only interested in the dominant mode. Thus **KDEm**, **KDE** and **RKDE** are robust to outliers compared with traditional numeric truth discovery models.

Results of experiments on `Population(outlier)` are showed in Table 5. Average time cost of each iteration in our **KDEm** model on this dataset is reported in Table 4. Similar to experiments on `Synthetic(unimodal)`, **KDEm** has the best performance and the performance of **KDE** can be improved by considering source quality. Also, traditional numeric models cannot estimate the truth precisely since they are too sensitive to outliers but results from **TruthFinder** and **AccuSim** are relatively good.

4.2 Multi-modality Detection and Anomaly Detection

A major feature of our **KDEm** model is that it can detect the controversy of the opinion distribution through multi-modality detection. For each entity, the number of reported representative values may indicate the number of modals of the opinion distribution. If this number is larger than one, the opinion of this entity may be controversial. Moreover, outliers can be naturally detected based on the estimated opinion distribution. Thus we can apply **KDEm** for anomaly detection. In this section, we conduct experiments on a set of synthetic datasets to verify the capability and superiority of our **KDEm** regarding multi-modality detection and anomaly detection on data from multiple sources. In addition, we provide a real world application, review rating summarization, to discover the controversy and consistency of users' feedback regarding products.

Datasets. A set of synthetic datasets `Synthetic(mix)` are used to verify the capability and superiority of **KDEm** and a set of real world datasets `Tripadvisor` [21,22] are used for the users' rating summarization.

- `Synthetic(mix)` is a set of one-dimensional ($d = 1$) synthetic datasets, whose generating procedure is similar to that of `Synthetic(unimodal)`. The major difference is that we generate 50 uni-modal entities and 50 bi-modal entities for each single dataset in `Synthetic(mix)`. Exactly the same generating procedure is applied on the 50 uni-modal entities. For the other 50 entities, we generate two representative values $t^*_{i1} = 1$ and $t^*_{i2} \sim N(1, 10)$. For entity N_i, source S_j randomly selects one of t^*_{i1} and t^*_{i2} and provides a claim x_{ij} from $N(t^*_{i1}, \sigma_j^2)$ or $N(t^*_{i2}, \sigma_j^2)$. Similarly we test our model for $p = 0.2$ and $\lambda = 10, 15, 20, 25, 30$ and generate 50 datasets for each pair of parameters. We also use the normalized z-score $(\{(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)/sd_i\}$ as input for our model and all the baseline methods.
- `Tripadvisor` is a set of review datasets and we only extract ratings from different users for different hotels. `Tripadvisor` dataset is originally published in [21, 22] and contains not only *overall* ratings but also aspect ratings regarding 1759 hotels. These aspects ratings include ratings regarding the *value*, *rooms*, *location*, *cleanliness*, *check in/front desk*, *service* and *business service*. Users may provide either ratings for all of these aspects or only a portion of them. We thus divide Tripadvisor dataset into eight subdatasets based on the overall rating and aspect ratings – `Tripadvisor(overall)`, `Tripadvisor(value)`, `Tripadvisor(rooms)`, `Tripadvisor(location)`, `Tripadvisor(cleanliness)`, `Tripadvisor(check in/front desk)`, `Tripadvisor(service)` and `Tripadvisor(business service)`. Basic statistics of `Tripadvisor` are included in Table 4.

Performance Measures. For each entity N_i in this kind of experiments, we may have multiple representative values of the opinion. Thus user's preference for this task should be "Discrete"+"Multiple" or "Continuous"+"Multiple". Notice that for multi-modality detection and anomaly detection, results based on these two kinds of preferences are the same because this task is only related to the clustering procedure. Since we only have groundtruth for `Synthetic(mix)`, we can only measure the performance on `Synthetic(mix)`. For `Tripadvisor`, we provide description analysis instead.

For experiments on `Synthetic(mix)`, if *thr* is a fixed parameter, we have

- $FPR = FP/(FP + TN)$;
- $TPR = TP/(TP + FN)$.

For multi-modality detection, suppose M is the number of modals we are interested in, K_i is the number of representative values reported from the model and K_i^* is the true number of representative values for the i-th entity. Then

- $TP = \sum_{i=1}^n \mathbf{1}\{K_i = K_i^* = M\}$
- $FP = \sum_{i=1}^n \mathbf{1}\{K_i = M, K_i^* \neq M\}$
- $FN = \sum_{i=1}^n \mathbf{1}\{K_i \neq M, K_i^* = M\}$
- $TN = \sum_{i=1}^n \mathbf{1}\{K_i \neq M, K_i^* \neq M\}$

Similarly, for anomaly detection, suppose

$$\hat{A}_{ij} = \begin{cases} 1, & \boldsymbol{x}_{ij} \text{ is detected as an anomaly observation;} \\ 0, & \text{otherwise.} \end{cases}$$

$$A_{ij} = \begin{cases} 1, & \boldsymbol{x}_{ij} \text{ is provided by an unreliable source } S_j; \\ 0, & \text{otherwise.} \end{cases}$$

Then we have

- $TP = \sum_{i=1}^n \sum_{j \in \mathcal{S}_i} \mathbf{1}\{\hat{A}_{ij} = A_{ij} = 1\}$;
- $FP = \sum_{i=1}^n \sum_{j \in \mathcal{S}_i} \mathbf{1}\{\hat{A}_{ij} = 1, A_{ij} = 0\}$;
- $FN = \sum_{i=1}^n \sum_{j \in \mathcal{S}_i} \mathbf{1}\{\hat{A}_{ij} = 0, A_{ij} = 1\}$;
- $TN = \sum_{i=1}^n \sum_{j \in \mathcal{S}_i} \mathbf{1}\{\hat{A}_{ij} = A_{ij} = 0\}$.

If we set the parameter *thr* to different values from 0 to 1, we can obtain a set of values $\{FPR_k, TPR_k, k = 1, 2, ..., k^*\}$ which are sorted based on FPR_k. Here we arbitrarily set $FPR_0 = TPR_0 = 0$ and $FPR_{k^*+1} = TPR_{k^*+1} = 1$ and an ROC curve can be obtained. Then we use the area under the ROC curve (AUC) to evaluate the performance:

$$\begin{aligned} AUC &= \int_{-\infty}^{\infty} TPR \; d(FPR) \\ &\approx \sum_k (TPR_k + TPR_{k-1})(FPR_k - FPR_{k-1})/2. \end{aligned} \tag{25}$$

Baseline. Similarly, we only introduce the baseline methods for experiments on `Synthetic(mix)`. For `Synthetic(mix)`, single truth existence cannot be ensured since a half of entities are bi-modal. Therefore, in addition to **KDEm**, we only apply **KDE** and **RKDE** on this kind of datasets and compare their multi-modality detection and anomaly detection

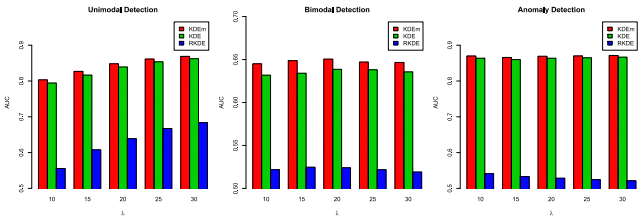


Figure 5: Results of experiments on synthetic mixed multi-modal datasets `Synthetic(mix)`.

Dataset	overall	value	rooms	location
# Bimodal entities	248	234	196	83
# Trimodal entities	1	5	1	1
Dataset	cleanliness	check in/ front desk	service	business service
# Bimodal entities	140	223	212	385
# Trimodal entities	5	9	7	10

Table 6: Number of detected multimodal entities in `Tripadvisor` datasets.

capabilities. Gaussian kernel is applied for these methods and $h_i = MAD_i^*$ for all entities in `Sythetic(mix)`.

Results. Results of experiments on the synthetic mixed multi-modal datasets `Synthetic(mix)` are showed in Figure 5. Based on Figure 5, for uni-modal, bi-modal, and anomaly detection, our model **KDEm** always has better performance than **KDE** and **RKDE** based on AUC. We also notice that **RKDE** has difficulty in distinguishing multi-modality and anomaly observations in this set of datasets. A possible reason could be that it tends to predict minority opinion instances as outliers when the number of claims is limited.

For `Tripadvisor`, since we don't have groundtruth, we only apply **KDEm** to estimate the trustworthy rating distribution and reliability scores of sources and use description analysis to evaluate the results. Since the values of claims for `Tripadvisor` can only be selected from $\{1, 2, 3, 4, 5\}$, we set a fixed bandwith $h_i = 0.8$ and $thr = 0.2$ for all the entities. Since the rating distributions of some entities in `Tripadvisor` may be multi-modal and the representative values need to be continuous, the user preference for **KDEm** on these datasets should be "Continuous"+"Multiple".

Average time costs of each iteration in **KDEm** on this set of datasets are reported in Table 4. The numbers of detected multi-modal entities in `Tripadvisor` are displayed in Table 6. From Table 6, we notice that the number of detected multi-modal entities in `Tripadvisor(location)` is much smaller while the number of detected multi-modal entities in `Tripadvisor(business service)` is larger than others. This indicates that users' opinions tend to agree on the location of a hotel while their feedbacks are diverse regarding the hotel service. In Figure 6, we provide histograms and estimated truth densities of one of uni-modal entities, one of bi-modal entities and one of tri-modal entities from `Tripadvisor(location)`.

We can obtain eight sets of source reliability scores and eight sets of number of predicted modals regarding different aspects respectively. For these two kinds of measures, the correlations between each pair of these eight datasets are calculated and displayed in Figure 7. From this figure, we notice that the correlations of source reliability scores between each pair of aspects are relatively strong while those of rating consistency are much weaker, which means source reliability tends to be consistent among different aspects while the consistency of claims tends to be independent of aspects.

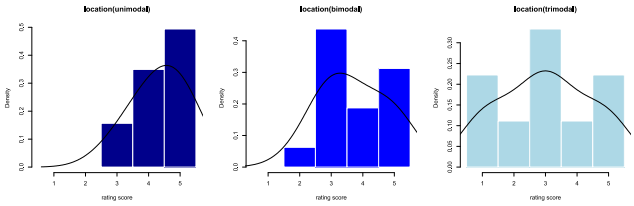
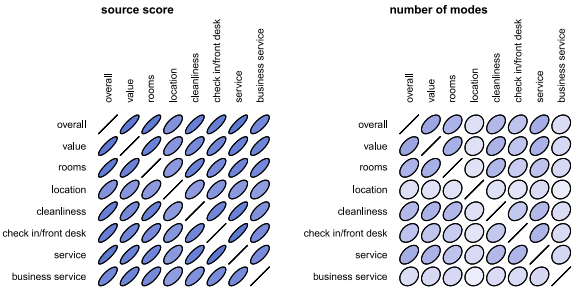


Figure 6: Histograms for examples of detected unimodal, bi-modal and tri-modal entity examples in the `Tripadvisor(location)` dataset.



- Darker ellipse indicates stronger correlation.
- For source reliability scores, the correlation is calculated based on sources which provide claims for both aspects of interest.

Figure 7: Pairwise correlation of source reliability scores and predicted numbers of modals for the `Tripadvisor` datasets.

5. RELATED WORK

The major technique in this study is inspired by kernel density estimation. Standard kernel density estimation (**KDE**) [14] is a non-parametric approach to estimate density function of a random variable. Since standard **KDE** may be sensitive to outliers, robust kernel density estimation (**RKDE**) [7], which is based on the idea of M-estimation, is proposed to overcome this limitation. However, the weight of each component function in **RKDE** is estimated based on a single entity rather than all the provided entities. Therefore, **RKDE** cannot estimate source reliability scores as precise as our **KDEm** model does.

Besides, various truth discovery models have been proposed to handle different scenarios [3–5, 9, 10, 13, 15–20, 23, 24, 26, 27] and these methods are summarized in a recent survey [12]. **TruthFinder** [23] is a Bayesian based iterative approach to estimate the truth and source reliability. The source consistency assumption in **TruthFinder** has been broadly applied in following-up studies. Then source dependency is considered in [3] and another model **AccuCopy** is proposed to solve this problem. As most truth discovery models, **TruthFinder** and **AccuCopy** are designed for categorical data but they both can be extended to handle numeric data by applying a similarity measure between claims. The extended version of **AccuCopy** is called **AccuSim**. In addition, particularly for the numeric data, a Bayesian framework **GTM** [25] is proposed to infer the real-valued truth and source reliability level. **TBP** [13] can be regarded as an extension of **GTM** to handle different difficulty levels of questions and to eliminate source bias. In [10], an

optimization framework **CRH** can be applied on heterogeneous data, where categorical and numeric data can be modeled together. It is noticed that most sources provide a few claims while only limited sources provide a number of claims. Thus in [9], this long-tail phenomenon is studied and a model **CATD** is proposed, in which the confidence interval of the source reliability is adopted to tackle this problem.

LTM [24] is a probabilistic graphical model where multiple values of truth are allowed. Notice that our uncertain-opinion assumption is different from this multiple truth assumption. In **LTM**, a reliable claim needs to include correct values and exclude wrong values as often as possible. However, in our study, trustworthy opinion is a random variable and multiple representative values can be summarized. A reliable claim can contain either one of these values. Fundamentally we do not estimate the "recall" of a source.

Apart from the truth discovery, some existing studies focused on the problem of statement truthfulness discovery. **T-verifyer** [11] is proposed to verify the truthfulness of fact statements. However, rather than finding out the truth from different claims, it determines whether a given statement is true by means of submitting the it to search engines.

Furthermore, this paper is much different from traditional opinion extraction and summarization task. Traditional opinion extraction and summarization is to extract informative words, summarize sentiments and associated degrees from given documents [8], where documents are treated independently and equally. In contrast, this paper focuses on identifying the trustworthiness of opinion, which is based on extracted informative opinion claims instead of raw documents. Specifically, our task is to find reliable opinion distribution from claims provided by multiple sources.

6. CONCLUSIONS

In this study, an uncertainty-aware model – **KDEm**, is introduced to estimate the probability density function of the trustworthy opinion from multiple sources. Based on the estimated distribution, representative opinion instances can be summarized as well. Experiments on synthetic and real-world datasets not only indicate that **KDEm** is more robust to extreme values claimed in multiple sources than traditional truth discovery models if the single truth existence can be ensured, but also shows that **KDEm** is good at detecting multi-modality and anomaly observations.

In the future, more loss functions and kernels can be theoretical studied to improve the accuracy and efficiency of **KDEm**. We only focus on quantitative information in this study but categorical data can be modeled by encoding them as high dimensional binary claims as well.

Acknowledgments

Research was sponsored in part by the U.S. Army Research Lab. under Cooperative Agreement No.W911NF-09-2-0053 (NSCTA), National Science Foundation IIS-1017362, IIS-1320617, IIS-1319973, IIS-1553411 and IIS-1354329, HDTRA1-10-1-0120, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), and UIUC. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies of the U.S. Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

7. REFERENCES

[1] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[2] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

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

[4] X. L. Dong, L. Berti-Equille, and D. Srivastava. Data fusion: resolving conflicts from multiple sources. In *WAIM*, 2013.

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

[6] A. Hinneburg and H.-H. Gabriel. Denclue 2.0: Fast clustering based on kernel density estimation. In *Advances in Intelligent Data Analysis VII*, pages 70–80. Springer, 2007.

[7] J. Kim and C. D. Scott. Robust kernel density estimation. *JMLR*, 13(1):2529–2565, 2012.

[8] L.-W. Ku, Y.-T. Liang, and H.-H. Chen. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 100107, 2006.

[9] 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.

[10] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *SIGMOD*, 2014.

[11] X. Li, W. Meng, and C. Yu. T-verifier: Verifying truthfulness of fact statements. In *ICDE*, 2011.

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

[13] R. W. Ouyang, L. Kaplan, P. Martin, A. Toniolo, M. Srivastava, and T. J. Norman. Debiasing crowdsourced quantitative characteristics in local businesses and services. In *IPSN*, 2015.

[14] E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, pages 1065–1076, 1962.

[15] J. Pasternack and R. Dan. Making better informed trust decisions with generalized fact-finding. In *IJCAI*, 2011.

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

[17] J. Pasternack and D. Roth. Latent credibility analysis. In *WWW*, 2013.

[18] G. J. Qi, C. C. Aggarwal, J. Han, and T. Huang. Mining collective intelligence in diverse groups. In *WWW*, 2013.

[19] V. G. V. Vydiswaran, C. X. Zhai, and D. Roth. Content-driven trust propagation framework. In *SIGKDD*, 2011.

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

[21] H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *SIGKDD*, 2010.

[22] H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis without aspect keyword supervision. In *SIGKDD*, 2011.

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

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

[25] 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.

[26] S. Zhi, B. Zhao, W. Tong, J. Gao, D. Yu, H. Ji, and J. Han. Modeling truth existence in truth discovery. In *SIGKDD*, 2011.

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