Resolving Conflicts in Heterogeneous Data by Truth Discovery and Source Reliability Estimation*

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

ABSTRACT

In many applications, one can obtain descriptions about the same objects or events from a variety of sources. As a result, this will inevitably lead to data or information conflicts. One important problem is to identify the true information (i.e., the *truths*) among conflicting sources of data. It is intuitive to trust reliable sources more when deriving the truths, but it is usually unknown which one is more reliable *a priori*. Moreover, each source possesses a variety of properties with different data types. An accurate estimation of source reliability has to be made by modeling multiple properties in a unified model. Existing conflict resolution work either does not conduct source reliability estimation, or models multiple properties separately. In this paper, we propose to resolve conflicts among multiple sources of heterogeneous data types. We model the problem using an optimization framework where truths and source reliability are defined as two sets of unknown variables. The objective is to minimize the overall weighted deviation between the truths and the multi-source observations where each source is weighted by its reliability. Different loss functions can be incorporated into this framework to recognize the characteristics of various data types, and efficient computation approaches are developed. Experiments on real-world weather, stock and flight data as well as simulated multi-source data demonstrate the necessity of jointly modeling different data types in the proposed framework[1].

1. INTRODUCTION

Recently, the Big Data challenge is motivated by a dramatic increase in our ability to extract and collect data from the physical world. One important property of Big Data is its wide *variety*, i.e., data about the same object can be obtained from various sources. For example, customer information can be found from multiple databases in a company, a patient's medical records may be scattered in different hospitals, and a natural event may be observed and recorded by multiple laboratories.

Due to recording or transmission errors, device malfunction, or malicious intent to manipulate the data, data sources usually contain noisy, outdated, missing or erroneous records, and thus multiple sources may provide conflicting information. In almost every industry, decisions based on untrustworthy information can cause serious damage. For example, erroneous account information in a company database may cause financial losses; wrong diagnosis based on incorrect measurements of a patient may lead to serious consequences; and scientific discoveries may be guided to the wrong direction if they are derived from incorrect data. Therefore, it is critical to *identify the most trustworthy answers from multiple sources of conflicting information*. This is a non-trivial problem due to the following two major challenges.

Source Reliability

Resolving conflicts from multiple sources have been studied in the database community for years [4, 5, 10, 13] resulting in multiple ways to handle conflicts in data integration. Among them, one commonly used approach to eliminate conflicts for categorical data is to conduct majority voting so that information with the highest number of occurrences is regarded as the correct answer; and for continuous values, we can simply take the mean or median as the answer. The issue of such Voting/Averaging approaches is that they assume all the sources are equally reliable, and thus the votes from different sources are uniformly weighted. In the complicated world that we have today, it is crucial to **estimate source reliability** to find out the correct information from conflicting data, especially when there exist sources providing low quality information, such as faulty sensors that keep emanating wrong data, and spam users who propagate false information on the Internet. However, there is no oracle telling us which source is more reliable and which piece of information is correct.

Heterogeneous Data

Motivated by the importance but lack of knowledge in source reliability, many truth discovery approaches have been proposed to estimate it and infer true facts without any supervision [3, 8, 9, 11, 12, 14–16, 18, 19, 21–24]. However, these approaches are mainly designed for single-type data and they do not take advantage of a joint inference on **data with heterogeneous types**.

In real data integration tasks, heterogeneous data is ubiquitous. An object usually possesses multiple types of data. For example, in the integration of multiple health record databases, a patient's record includes age, height, weight, address, measurements, etc; we may want to infer correct information for a city's population, area, mayor, and founding year among conflicting information presented

*The first two authors contributed equally to this work and should be considered as joint first authors.

[1] Codes and data sets can be downloaded from `http://www.cse.buffalo.edu/~jing/crh.htm`

SIGMOD'14, June 22–27, 2014, Snowbird, UT, USA.
Copyright 2014 ACM 978-1-4503-2376-5/14/06 ...$15.00.
http://dx.doi.org/10.1145/2588555.2610509 .

on the Internet; and when we combine the predictions from multiple weather forecast tools, we need to resolve conflicts in weather conditions, temperature, humidity, wind speed, wind direction, etc. In all these cases, the data to be integrated involve categorical, continuous or even more complicated data types.

Due to the wide existence of missing values, we usually do not have sufficient amount of data to estimate source reliability correctly purely from one type of data. When source reliability is consistent on the entire data set, which is often valid in reality, a model that infers from various data types together will generate accurate estimates of source reliability, which will in turn help infer accurate information. Therefore, instead of separately inferring trustworthy information for individual data types, we should develop a unified model that conducts a joint estimation on all types of data simultaneously.

However, it is non-trivial to unify different types of data in one model. During source reliability estimation, we need to estimate how close a source input is to the correct answer, but different data types should be treated differently in this process because the concept of closeness varies among different data types. For categorical data, each observation will be either correct or wrong (i.e., whether the observation is the same as or different from the true fact). It is very different when a property has continuous values. For example, if the true temperature is 80F, then an observation of 79F is closer to the true value than 70F. If such differences are not taken into account and we regard each continuous input as a fact, we will inevitably make wrong estimates of source reliability and derive incorrect results. Therefore, we need a framework that can seamlessly integrate data of heterogeneous data types by estimating information trustworthiness.

Summary of Proposed CRH Framework

These observations motivate us to develop a ***C**onflict **R**esolution on **H**eterogeneous Data* (CRH) framework to infer the truths (also referred to as the true information or correct answers) from multiple conflicting sources each of which involves a variety of data types. We formulate the problem as an optimization problem to minimize the overall weighted deviation between the identified truths and the input. The weights in the objective function correspond to source reliability degrees. We propose to leverage heterogeneous data types by allowing any loss function for any type of data, and find out both truths and source reliability by solving the joint optimization problem. In the experiments (Section 3), we show that the proposed CRH framework outperforms existing conflict resolution approaches applied separately or jointly on heterogeneous data because each baseline approach either does not conduct source reliability estimation, or takes incomplete single-type data, or ignores the unique characteristics of each data type.

In summary, we make the following contributions:

- We design a general optimization framework to model the conflict resolution problem on heterogeneous data by incorporating source reliability estimation. The proposed objective function characterizes the overall difference between unknown truths and input data while modeling source reliability as unknown source weights in the framework.
- Under this framework, weight assignment schemes are introduced to capture source reliability distributions. Various loss functions can be plugged into the framework to characterize different types of data. In particular, we discuss several common choices and illustrate their effectiveness in modeling conflict resolution on heterogeneous data.
- We propose an algorithm to solve the optimization problem by iteratively updating truths and source weights. We derive effective solutions for commonly used loss functions and weight assignment schemes, and show the convergence of the algorithm. The running time is linear in the number of observations. Also, the proposed approach can be adapted to MapReduce model, so it can scale to very large data sets.
- We validate the proposed algorithm on both real-world and simulated data sets, and the results demonstrate the advantages of the proposed approach in resolving conflicts from multi-source heterogeneous data. The CRH framework can improve the performance of existing approaches due to its ability of tightly coupling various data types in the conflict resolution and source reliability estimation process. Running time on both single and Hadoop cluster machines demonstrates its efficiency.

2. METHODOLOGY

In this section, we describe our design of the CRH model, which computes truths and source weights from multi-source heterogeneous data. We formulate the conflict resolution problem as an optimization problem which models the truths as the weighted combination of the observations from multiple sources and incorporates a variety of loss functions for heterogeneous data types. An iterative weight and truth computation procedure is introduced to solve this optimization problem. Under this general framework, we present several loss functions and constraints, discuss other possible choices and analyze the time complexity.

2.1 Problem Formulation

We start by introducing important terms and defining the conflict resolution problem. We use an example on demographic database (Table 1) to illustrate these concepts.

DEFINITION 1. *An* object *is a person or thing of interest; a* property *is a feature used to describe the object; and a* source *describes the place where information about objects' properties can be collected.*

EXAMPLE 1. *"Bob" is an object; height is a property; and a database that provides the information is a source.*

DEFINITION 2. *An* observation *is the data describing a property of an object from a source.*

EXAMPLE 2. *The observation on Bob's height from Source 1 is 1.74m.*

DEFINITION 3. *An* entry *is a property of an object, and the* truth *of an entry is defined as its accurate information, which is unique.*

EXAMPLE 3. *Bob's height is an entry. The real height of Bob is the truth of the entry.*

The mathematical notation is as follows. Suppose there are N objects, each of which has M properties whose data types can be different, and these objects are observed by K sources.

Input

DEFINITION 4. *The observation of the m-th property for the i-th object made by the k-th source is $v_{im}^{(k)}$.*

DEFINITION 5 (SOURCE OBSERVATION TABLES).
The k-th source $\mathcal{X}^{(k)}$ is the collection of observations made on all the objects by the k-th source. It is denoted as a matrix whose im-th entry is $v_{im}^{(k)}$. $\{\mathcal{X}^{(1)}, \mathcal{X}^{(2)}, \ldots, \mathcal{X}^{(K)}\}$ are the K source observation tables.

Table 1: Observation Tables

	$\mathcal{X}^{(1)}$		$\mathcal{X}^{(2)}$		$\mathcal{X}^{(3)}$	
Object	City	Height	City	Height	City	Height
Bob	NYC	1.72	NYC	1.70	NYC	1.90
Mary	LA	1.62	LA	1.61	LA	1.85
Kate	NYC	1.74	NYC	1.72	LA	1.65
Mike	NYC	1.72	LA	1.70	DC	1.85
Joe	DC	1.72	NYC	1.71	NYC	1.85

Table 2: Ground Truth and Conflict Resolution Results

	Ground Truth		Voting/Averaging		CRH	
Object	City	Height	City	Height	City	Height
Bob	NYC	1.72	NYC	1.77	NYC	1.72
Mary	LA	1.62	LA	1.69	LA	1.62
Kate	NYC	1.75	NYC	1.70	NYC	1.74
Mike	NYC	1.71	DC	1.76	NYC	1.72
Joe	DC	1.73	NYC	1.76	DC	1.72

EXAMPLE 4. *Suppose we have three databases (sources) that provide the information on a group of people. Table 1 shows the observation tables of $\mathcal{X}^{(1)}$, $\mathcal{X}^{(2)}$, and $\mathcal{X}^{(3)}$, which store all observations made by Sources 1, 2 and 3 respectively. Source 1 states that the citizenship for Bob (row 1 and column 1 in $\mathcal{X}^{(1)}$) is "NYC", so $v_{11}^{(1)}$ is "NYC".*

Remark. To simplify the notations, we assume that the observations of all the sources about all the objects are available in the formulation. However, the proposed framework is general enough to cover the cases with missing observations. More discussion can be found in Section 2.5.

Output of CRH Framework

DEFINITION 6 (TRUTH TABLE).
$v_{im}^{(*)}$ denotes the truth of the m-th property for the i-th object. The truths of all the objects on all the properties are stored in a truth table $\mathcal{X}^{(*)}$ whose im-th entry is $v_{im}^{(*)}$.

EXAMPLE 5. *Table 2 shows the ground truth, the results obtained by Voting/Averaging, and the results obtained by the proposed framework CRH from the input tables. CRH resolves conflicts from different sources for each entry. It provides more accurate results comparing with Voting/Averaging.*

Remark. Comparing with the ground truths, it is clear that Source 1 provides accurate information more often (more reliable) while Source 2 and 3 are not very reliable. Due to the various reliability degrees, voting cannot work well. For example, voting fails on Joe's citizenship because the majority answer is stated by the unreliable sources and majority voting gives a wrong answer – "NYC". Although only Source 1 states that Joe's citizenship is "DC", as it is a reliable source, we should take this observation as the truth.

However, ground truths and source reliability are usually unknown *a priori*. Existing truth discovery approaches [3, 8, 9, 11, 12, 14–16, 18, 21–24] and the proposed method try to estimate source reliability in an unsupervised manner. The difference from existing approaches is that the proposed approach estimates source reliability jointly with heterogeneous data, which is discussed as follows.

DEFINITION 7 (SOURCE WEIGHTS).
Source weights are denoted as $\mathcal{W} = \{w_1, w_2, \ldots, w_K\}$ in which w_k is the reliability degree of the k-th source. A higher w_k indicates that the k-th source is more reliable and observations from this source is more likely to be accurate.

Remark. As for the approaches that incorporate source reliability estimation, if this estimation is only conducted on individual properties separately, the estimated reliability result is not accurate enough due to insufficient observations. In the example shown in Table 1, if we only consider citizenship, the reliability degrees of Sources 1 and 3 cannot be easily distinguished. In contrast, if we consider all the properties in the source reliability estimation, we can find that Source 1 is better than Source 3 in terms of the ability of providing accurate information. To characterize this phenomenon, the proposed framework unifies heterogeneous properties in the source reliability estimation. It will output both source weights and a truth table which are computed simultaneously by estimating source reliability from all the properties.

2.2 CRH Framework

The basic idea behind the proposed framework is that reliable sources provide trustworthy observations, so the truths should be close to the observations from reliable sources, and thus we should minimize the weighted deviation from the truths to the multi-source input where the weight reflects the reliability degree of sources. Based on this principle, we propose the following optimization framework that can unify heterogeneous properties in this process:

$$\min_{\mathcal{X}^{(*)},\mathcal{W}} \quad f(\mathcal{X}^{(*)}, \mathcal{W}) = \sum_{k=1}^{K} w_k \sum_{i=1}^{N} \sum_{m=1}^{M} d_m(v_{im}^{(*)}, v_{im}^{(k)})$$
$$\text{s.t.} \quad \delta(\mathcal{W}) = 1, \quad \mathcal{W} \in \mathcal{S}. \qquad (1)$$

We are trying to search for the values for two sets of unknown variables $\mathcal{X}^{(*)}$ and $\mathcal{W}$, which correspond to the collection of truths and source weights respectively, by minimizing the objective function $f(\mathcal{X}^{(*)}, \mathcal{W})$. There are two types of functions that need to be plugged into this framework:

- *Loss function.* d_m refers to a loss function defined based on the data type of the m-th property. This function measures the distance between the truth $v_{im}^{(*)}$ and the observation $v_{im}^{(k)}$. This loss function should output a high value when the observation deviates from the truth and a low value when the observation is close to the truth.
- *Regularization function.* $\delta(\mathcal{W})$ reflects the distributions of source weights. It is also required mathematically. If each source weight w_k is unconstrained, then the optimization problem is unbounded because we can simply take w_k to be $-\infty$. To constrain the source weights $\mathcal{W}$ into a certain range, we need to specify the regularization function $\delta(\mathcal{W})$ and the domain $\mathcal{S}$. Note that we set the value of $\delta(\mathcal{W})$ to be 1 for the sake of simplicity. Different constants for $\delta(\mathcal{W})$ do not affect the results, as we can divide $\delta(\mathcal{W})$ by the constant.

These two types of functions should be chosen based on our knowledge on the characteristics of heterogeneous data and the source reliability distributions, and more details about these functions will be discussed later. Intuitively, if a source is more reliable (i.e., w_k is high), high penalty will be received if this source's observation is quite different from the truth (i.e., difference between $v_{im}^{(*)}$ and $v_{im}^{(k)}$ is big). In contrast, the observation made by an unreliable source with a low w_k is allowed to be different from the truth. In order to minimize the objective function, the truths $\mathcal{X}^{(*)}$ will rely more on the sources with high weights.

The truths $\mathcal{X}^{(*)}$ and source weights $\mathcal{W}$ should be learned together by optimizing the objective function through a joint procedure. In an optimization problem that involves two sets of variables, it is natural to iteratively update the values of one set to minimize the objective function while maintaining the values of another set

```
Algorithm 1 CRH Framework
Input: Data from $K$ sources: $\{\mathcal{X}^{(1)}, \dots, \mathcal{X}^{(K)}\}$.
Output: Truths $\mathcal{X}^{(*)} = \{v_{im}^{(*)}\}_{i=1,m=1}^{N,M}$, source weights $\mathcal{W} = \{w_1, \dots, w_K\}$.

Initialize the truths $\mathcal{X}^{(*)}$;
repeat
    Update source weights $\mathcal{W}$ according to Eq(2) to reflect sources' reliability
    based on the estimated truths;
    for $i \leftarrow 1$ to $N$ do
        for $m \leftarrow 1$ to $M$ do
            Update the truth of the $i$-th object on the $m$-th property $v_{im}^{(*)}$ according
            to Eq(3) based on the current estimation of source weights;
        end for
    end for
until Convergence criterion is satisfied;
return $\mathcal{X}^{(*)}$ and $\mathcal{W}$.
```

until convergence. This iterative two-step procedure, referred to as block coordinate descent approach [2], will keep reducing the value of the objective function. To minimize the objective function in Eq(1), we iteratively conduct the following two steps.

Step I: Source Weights Update. With an initial estimate of the truths $\mathcal{X}^{(*)}$, we first weight each source based on the difference between the truths and the observations made by the source:

$$\mathcal{W} \leftarrow \arg\min_{\mathcal{W}} f(\mathcal{X}^{(*)}, \mathcal{W}) \quad \text{s.t.} \quad \delta(\mathcal{W}) = 1, \quad \mathcal{W} \in \mathcal{S}. \tag{2}$$

At this step, we fix the values for the truths and compute the source weights that jointly minimize the objective function subject to the regularization constraints.

Step II: Truths Update. At this step, the weight of each source w_k is fixed, and we update the truth for each entry to minimize the difference between the truth and the sources' observations where sources are weighted by their reliability degrees:

$$v_{im}^{(*)} \leftarrow \arg\min_{v} \sum_{k=1}^{K} w_k \cdot d_m(v, v_{im}^{(k)}). \tag{3}$$

By deriving the truth using this equation for every entry, we can obtain the collection of truths $\mathcal{X}^{(*)}$ which minimizes $f(\mathcal{X}^{(*)}, \mathcal{W})$ with fixed $\mathcal{W}$.

The pseudo code of this framework is summarized in Algorithm 1. We start with an initial estimate of truths and then iteratively conduct the source weight update and truth update steps until convergence. In the following, we explain the two steps in detail using example functions, and discuss the convergence and other practical issues of the proposed approach.

2.3 Source Weight Assignment

We propose the following regularization function:

$$\delta(\mathcal{W}) = \sum_{k=1}^{K} \exp(-w_k). \tag{4}$$

This function regularizes the value of w_k by constraining the sum of $\exp(-w_k)$.

THEOREM 1. *Suppose that the truths are fixed, the optimization problem Eq(1) with constraint Eq(4) is convex. Furthermore, the global optimal solution is given by*

$$w_k = -\log\left(\frac{\sum_{i=1}^{N}\sum_{m=1}^{M} d_m(v_{im}^{(*)}, v_{im}^{(k)})}{\sum_{k'=1}^{K}\sum_{i=1}^{N}\sum_{m=1}^{M} d_m(v_{im}^{(*)}, v_{im}^{(k')})}\right). \tag{5}$$

PROOF. Since the truths are fixed, the optimization problem Eq(1) has only one set of variables $\mathcal{W}$. To prove the convexity of Eq(1), we introduce another variable t_k so that $t_k = \exp(-w_k)$. Now we express the optimization problem in terms of t_k:

$$\begin{aligned} \min_{\{t_k\}_{k=1}^{K}} \quad & f(t_k) = \sum_{k=1}^{K} -\log(t_k) \sum_{i=1}^{N}\sum_{m=1}^{M} d_m(v_{im}^{(*)}, v_{im}^{(k)}) \\ & \text{s.t.} \quad \sum_{k=1}^{K} t_k = 1. \end{aligned} \tag{6}$$

The constraint in Eq(6) is linear in t_k, which is affine. The objective function is a linear combination of negative logarithm functions and thus it is convex. Therefore, the optimization problem Eq(1) with Eq(4) is convex, and any local optimum is also global optimum [6].

We use the method of Lagrange multipliers to solve this optimization problem. The Lagrangian of Eq(6) is given as:

$$\begin{aligned} L(\{t_k\}_{k=1}^{K}, \lambda) \quad = \quad & \sum_{k=1}^{K} -\log(t_k) \sum_{i=1}^{N}\sum_{m=1}^{M} d_m(v_{im}^{(*)}, v_{im}^{(k)}) \\ & +\lambda(\sum_{k=1}^{K} t_k - 1), \end{aligned} \tag{7}$$

where λ is a Lagrange multiplier. Let the partial derivative of Lagrangian with respect to t_k be 0, and we can get:

$$\sum_{i=1}^{N}\sum_{m=1}^{M} d_m(v_{im}^{(*)}, v_{im}^{(k)}) = \lambda t_k. \tag{8}$$

From the constraint that $\sum_{k=1}^{K} t_k = 1$, we can derive that

$$\lambda = \sum_{k'=1}^{K}\sum_{i=1}^{N}\sum_{m=1}^{M} d_m(v_{im}^{(*)}, v_{im}^{(k')}). \tag{9}$$

Plugging Eq(9) and $w_k = -\log(t_k)$ into Eq(8), we obtain Eq(5). □

This weight computation equation indicates that a source's weight is inversely proportional to the difference between its observations and the truths at the log scale. The negative log function maps a number in the range of 0 and 1 to a range of 0 and ∞, so it helps to enlarge the difference in the source weights. A source whose observations are more often close to the truths will have a higher weight. Therefore, Eq(4) is a reasonable constraint function which leads to the meaningful and intuitive weight update formula.

In order to distinguish the source weights even better so that reliable sources can play a more important role in deriving the truths, we use the maximum rather than the sum of the deviations as the normalization factor when computing the weights. It still ensures that a source's weight is inversely proportional to the difference between its observations and the truths at the log scale.

EXAMPLE 6. *Considering the databases given in Table 1, if* Voting/Averaging *results are adopted as the initial truths, we can get weights 0.67, 0.41, and 0.01 for Sources 1, 2, and 3 respectively using the proposed weight computation equation (Eq(5)). As the proposed method is iterative, we show the converged source weights here (more details are given in Table 3 after we discuss all the steps in the iterative framework). Source 1 is the most reliable one and Source 3 is the least reliable one, as Source 1 makes the fewest errors while Source 3 makes the most errors.*

The aforementioned weight assignment scheme considers a combination of sources. By setting different regularization functions, we can conduct source selection under the framework. For example, the following function defined based on L^p-norm can be used to select sources:

$$\begin{aligned}\delta(\mathcal{W}) = &\sqrt[p]{w_1^p + w_2^p + \ldots + w_K^p} = 1,\\ &w_k \in \mathbb{R}^+ \ (k = 1, \ldots, K),\end{aligned} \tag{10}$$

where p is a positive integer. When p equals to 1 or 2, it corresponds to the most widely used L^1-norm or L^2-norm. If L^p-norm regularization is employed, the optimal value of the problem in Eq(1) will be 0, which is achieved when we select one of the sources and set its weight to be 1, set all the other source weights to be 0, and simply regard the chosen source's observations as the truths. Different from the regularization function shown in Eq(4), this regularization function does not combine multiple sources but rather assumes that there only exists one reliable source.

We can also incorporate integer constraints to conduct source selection with more than one source, i.e., choose j sources out of all K sources:

$$\begin{aligned}\delta(\mathcal{W}) = &\frac{1}{j}(w_1 + w_2 + \ldots + w_K) = 1,\\ &w_k \in \{0, 1\} \ (k = 1, \ldots, K).\end{aligned} \tag{11}$$

If $w_k = 1$, the k-th source is selected in truth computation, otherwise its observations will be ignored when updating the truths in the next step. Due to the integer constraints defined in Eq(11), Eq(1) becomes an NP-hard problem. Approximation algorithms can be developed to solve this problem, the details of which are omitted.

In many problems, we will benefit from integrating the observations from multiple sources, but there is a variation in the overall reliability degrees. Therefore, in this paper, we focus on the weight assignment scheme with max normalization factor where sources are integrated and variation is emphasized.

2.4 Truth Computation

The truth computation step (Eq(3)) depends on the data type and loss function. We respect the characteristics of each data type and utilize different loss functions to describe different notions of deviation from the truths for different data types. Accordingly, truth computation will differ among various data types. Below we discuss truth computation in detail based on several loss functions for categorical and continuous data, the two most common data types.

On categorical data, the most commonly used loss function is 0-1 loss in which an error is incurred if the observation is different from the truth. Formally, if the m-th property is categorical, the deviation from the truth $v_{im}^{(*)}$ to the observation $v_{im}^{(k)}$ is defined as:

$$d_m(v_{im}^{(*)}, v_{im}^{(k)}) = \begin{cases} 1 & \text{if } v_{im}^{(k)} \neq v_{im}^{(*)}, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

THEOREM 2. *Suppose that the weights are fixed, based on 0-1 loss function, to minimize the objective function at this step (Eq(3)), the truth on the m-th property of the i-th object should be the value that receives the highest weighted votes among all possible values:*

$$v_{im}^{(*)} \leftarrow \arg\max_{v} \sum_{k=1}^{K} w_k \cdot \mathbb{1}(v, v_{im}^{(k)}), \tag{13}$$

where $\mathbb{1}(x, y) = 1$ if $x = y$, and 0 otherwise.

PROOF. We plug Eq(12) into the objective function in Eq(1):

$$v_{im}^{(*)} \leftarrow \arg\min_{v} \sum_{k=1}^{K} w_k \cdot d_m(v, v_{im}^{(k)}), \tag{14}$$

which is equivalent to Eq(13). □

This computation follows the principle that an observation stated by reliable sources will be regarded as the truth.

EXAMPLE 7. *If we use the weights 0.67, 0.41, and 0.01 for Sources 1, 2, and 3 from Example 6, the proposed approach will output the citizenship for Joe as "DC" (different from "NYC" output by voting), because Source 1's weight is higher than the sum of Source 2 and 3's weights. With an accurate source weight estimation, we correct the mistake given by Voting/Averaging.*

For the scenarios where multiple values of $v_{im}^{(*)}$ are probable, we introduce a strategy to incorporate probability into truth computation. This strategy is probabilistic-based and we assume that observations from reliable sources should have higher probability to be true. We represent categorical data by binary index vectors, which characterize the probability distributions of observations over all possible values. Formally, if the m-th property has L_m possible values and $v_{im}^{(k)}$ is the l-th value, then the index vector $I_{im}^{(k)}$ for $v_{im}^{(k)}$ is defined as:

$$I_{im}^{(k)} = (0, \ldots, \overset{l}{1}, 0, \ldots, 0)^T. \tag{15}$$

We can use squared loss function to describe the distance between the observation index vector $I_{im}^{(k)}$ and the truth vector $I_{im}^{(*)}$:

$$\begin{aligned} d_m(v_{im}^{(*)}, v_{im}^{(k)}) &= d_m(I_{im}^{(*)}, I_{im}^{(k)}) \\ &= (I_{im}^{(*)} - I_{im}^{(k)})^T (I_{im}^{(*)} - I_{im}^{(k)}). \end{aligned} \tag{16}$$

THEOREM 3. *Suppose that the weights are fixed, the optimization problem Eq(1) with Eq(16) is convex. The optimal $I_{im}^{(*)}$ should be the weighted mean of the probability vectors of all the sources:*

$$I_{im}^{(*)} \leftarrow \frac{\sum_{k=1}^{K} w_k \cdot I_{im}^{(k)}}{\sum_{k=1}^{K} w_k}. \tag{17}$$

PROOF. Since the weights are fixed, the optimization problem Eq(1) has only one set of variables $\mathcal{X}$ and the optimization problem is unconstrained. As $d_m(I_{im}^{(*)}, I_{im}^{(k)})$ is convex, Eq(1) is a linear combination of convex functions, and thus it is convex.

We plug Eq(16) into the objective function in Eq(1) and then let the partial derivative with respect to $I_{im}^{(*)}$ be 0:

$$\sum_{k=1}^{K} w_k \sum_{i=1}^{N} \sum_{m=1}^{M} 2(I_{im}^{(*)} - I_{im}^{(k)})^T = 0, \tag{18}$$

which leads to Eq(17). □

Here, $I_{im}^{(*)}$ denotes the probability distribution of the truths, in which $v_{im}^{(*)}$ is the corresponding value with the largest probability in $I_{im}^{(*)}$, i.e., the most possible value.

EXAMPLE 8. *We use the same scenario as in Example 7. Source weights are 0.67, 0.41, and 0.01; the observations are "DC", "NYC", and "NYC" from the corresponding sources. If "DC" is coded as (1,0) and "NYC" as (0,1), we can calculate the truth vector for that entry as $I_{im}^{(*)} = \frac{(1,0)\times 0.67+(0,1)\times 0.41+(0,1)\times 0.01}{0.67+0.41+0.01} = (0.61, 0.39)$. Therefore, the truth is "DC" with a probability of 0.61.*

Comparing with the 0-1 loss strategy, this strategy gives a soft decision instead of a hard decision. However, this method has relatively high space complexity due to the representation of categories for input data.

As for the continuous data, the loss function should characterize the distance from the input to the truth with respect to the variance of entries across sources. One common loss function is the normalized squared loss, which is defined as:

$$d_m(v_{im}^{(*)}, v_{im}^{(k)}) = \frac{(v_{im}^{(*)} - v_{im}^{(k)})^2}{std(v_{im}^{(1)}, \dots, v_{im}^{(K)})}. \tag{19}$$

THEOREM 4. *Suppose that the weights are fixed, the optimization problem Eq(1) with Eq(19) is convex. The truth that minimizes the overall weighted distance should be the weighted average of the observations:*

$$v_{im}^{(*)} \leftarrow \frac{\sum_{k=1}^K w_k \cdot v_{im}^{(k)}}{\sum_{k=1}^K w_k}. \tag{20}$$

PROOF. The proof of convexity is similar to the proof in Theorem 3. We plug Eq(19) into the objective function Eq(1) and then let the partial derivative with respect to $v_{im}^{(*)}$ be 0:

$$\sum_{k=1}^K w_k \sum_{i=1}^N \sum_{m=1}^M 2(v_{im}^{(*)} - v_{im}^{(k)})/std(v_{im}^{(1)}, \dots, v_{im}^{(K)}) = 0. \tag{21}$$

Therefore, we can get the optimal truth shown in Eq(20). □

EXAMPLE 9. *Consider Bob's height from Table 1. The observations from Sources 1, 2, and 3 are 1.72, 1.70, and 1.90 respectively. Suppose the weights are 0.67, 0.41, and 0.01 for Sources 1, 2, and 3 respectively. Then Bob's height can be calculated as:* $\frac{1.72\times0.67+1.70\times0.41+1.90\times0.01}{0.67+0.41+0.01} = 1.71$. *It is obvious that the result is closer towards the observation from Source 1 because it is more reliable.*

This truth computation strategy simulates the idea that observations from a reliable source should contribute more to the computation of the truth. However, this method is sensitive to the existence of outliers, as an outlier will receive a huge loss because of the square term, and thus can only work well in the data set in which outliers are removed.

To mitigate the effect of outliers, we can use the normalized absolute deviation as the loss function on continuous data:

$$d_m(v_{im}^{(*)}, v_{im}^{(k)}) = \frac{|v_{im}^{(*)} - v_{im}^{(k)}|}{std(v_{im}^{(1)}, \dots, v_{im}^{(K)})}. \tag{22}$$

THEOREM 5. *Based on Eq(22), the truth that minimizes the overall weighted absolute deviation should be the weighted median.*

Proof is omitted as the derivation is similar to the previous cases. Specifically, we use the following definition of weighted median[2]. Given a set of numbers $\{v^1, ..., v^K\}$ with weights $\{w_1, ...w_K\}$, the weighted median of this set is the number v^j, such that

$$\sum_{k:v^k<v^j} w_k < \frac{1}{2}\sum_{k=1}^K w_k \quad \& \quad \sum_{k:v^k>v^j} w_k \leqslant \frac{1}{2}\sum_{k=1}^K w_k. \tag{23}$$

The sum of weights on the numbers that are smaller than the weighted median, and the sum of weights on the numbers that are greater than the weighted median should both be roughly half of the total weights on the whole set. To find the weighted median, we compare the cumulative sum computed on numbers smaller than v^j or greater than v^j. Note that conventional median can be regarded as a special case where we give the same weight to all the numbers so that median becomes the number separating the higher half from the lower half. It is known that median is less sensitive to the existence of outliers, and thus the weighted median approach for truth computation is more desirable in noisy environments.

EXAMPLE 10. *Using the same example as in Example 9, we first sort all observations in ascending order: (1.70, 1.72, 1.90), and the corresponding source weight is (0.41, 0.67, 0.01). Then Bob's height by weighted median using Eq(23) is 1.72.*

Besides the aforementioned loss functions, the proposed general framework can take any loss function that is selected based on data types and distributions. Some other examples include Mahalanobis distance for continuous data, edit distance or KL divergence for text data, etc. To deal with complex data types, we can either use loss functions defined on raw data or on abstraction of raw data, such as motifs in time series, frequent sub-graphs in graphs, and segments in images. The framework can even be adapted to take the ensemble of multiple loss functions for a more robust loss computation. We can also convert a similarity function into a loss function, which allows the usage of numerous techniques in similarity computation developed in the data integration community.

2.5 Discussions & Practical Issues

Here we first discuss several important issues to make the framework practical including *initialization*, *convergence*, *normalization*, and *missing values*. Then we show the algorithm flow using a running example. Finally, we analyze the time complexity of the proposed CRH framework.

Initialization. The initialization of the truths can be obtained using existing conflict resolution approaches. In our experiments, we find that the result from Voting/Averaging approaches is typically a good start. In fact, initialization will not affect the final results if the optimization problem is convex.

Convexity and Convergence. The convexity depends on the loss functions and regularization function. An example of a family of convex loss functions is *Bregman divergence* [1], which includes a variety of loss functions such as squared loss, logistic loss, Itakura-Saito distance, squared Euclidean distance, Mahalanobis distance, KL-divergence and generalized I-divergence. Using several loss functions discussed in this paper, we prove the convergence of the CRH framework as follows. Note that the proof on the convergence can be derived in the same way if other convex functions are used.

THEOREM 6. *When Eq(4) is used as constraint, Eq(16) or/and Eq(19) is/are used as loss functions, the convergence of CRH framework is guaranteed.*

PROOF. For the optimization problem of Eq(1), we proved earlier in Theorem 1 that the unique minimum with respect to weights $\mathcal{W}$ can be achieved when truths $\mathcal{X}^{(*)}$ are fixed; we proved in Theorem 3 and Theorem 4 that the unique minimum with respect to truths $\mathcal{X}^{(*)}$ can be achieved when weights $\mathcal{W}$ are fixed. According to the proposition on the convergence of block coordinate descent [2], we can derive that the proposed iterative procedure will converge to a stationary point of the optimization problem. □

In Algorithm 1, the convergence criterion is that the decrease in the objective function is small enough compared with the previous iterations. In the experiments we find that the convergence of this approach is easy to judge because the first several iterations incur a huge decrease in the objective function, and once it converges, the results become stable. Experimental results on convergence are shown in Section 3.2.3.

Although the analysis on non-convex or non-differentiable functions need to be conducted differently [17,20], we find that some of these approaches work well in practice, such as the absolute deviation for continuous data.

Normalization. Another important issue is the normalization of deviations on each property. As illustrated in the weight computation equation (Eq(5)), we need to sum up the deviations to the truths

[2] http://www.mathworks.com/matlabcentral/fileexchange /23077-weighted-median/content/weightedMedian.m

across different properties. If various loss functions applied on different properties have significantly different scales, the weight computation will be biased towards the property that has bigger range in the deviation. To solve this issue, we normalize the output of each loss function on each property so that the deviation computed on all the properties fall into the same range.

Missing Values. Note that for the sake of simplicity, we assume that all the sources observe all the objects on all the properties in the proposed optimization framework (Eq(1)), but it can be easily modified to handle missing values when different sources observe different subsets of the objects on different subsets of properties. When the number of observations made by different sources is quite different, we can normalize the overall distance of each source by the number of observations.

Running Example. Using the same databases in Table 1, we display both the truth tables and the source weights for the first three iterations using Voting/Averaging as initialization. We use weighted vote for categorical data and weighted median for continuous data to update truths. The algorithm updates source weights and truths repeatedly until the convergence criterion is satisfied. The results are shown in Table 3 and 4. Comparing with the ground truths in Table 2, the truth table gets more and more accurate with more iterations. The algorithm converges after three iterations.

Table 3: Source Weights for the First Three Iterations

	Iteration 1	Iteration 2	Iteration 3
Source 1	0.2729	0.5552	0.6734
Source 2	0.2521	0.4539	0.4077
Source 3	0.1349	0.0149	0.0141

Table 4: Truth Tables for the First Three Iterations

	Iteration 1		Iteration 2		Iteration 3	
Object	City	Height	City	Height	City	Height
Bob	NYC	1.72	NYC	1.72	NYC	1.72
Mary	LA	1.62	LA	1.62	LA	1.62
Kate	NYC	1.72	NYC	1.74	NYC	1.74
Mike	NYC	1.72	NYC	1.72	NYC	1.72
Joe	NYC	1.72	DC	1.72	DC	1.72

Time Complexity. When we utilize the aforementioned functions Eq(5), Eq(13) and Eq(23), the running time is linear with respect to the total number of observations, i.e., $O(KNM)$, where K is the number of sources, N is the number of objects, and M is the number of properties. In addition, the algorithm can be adapted to run on the MapReduce platform to further speed up the process on large-scale data sets. It is obvious that the truth computation step (Eq(3)) can be executed independently for each object and thus this step is easy to parallelize. If we use Eq(5) to compute source weights, this step can be expressed using summation form [7], and thus it can be parallelized by aggregating partial sums. The time complexity is experimentally validated in the following section, in which we show the running time of the proposed approach on both single machine and Hadoop cluster.

2.6 Summary

Our major contribution is that we unify data of various types in truth discovery to resolve conflicts on heterogeneous data. The proposed optimization framework (Eq(1)), which targets at minimizing overall weighted difference between truths and input data, provides a nice way to combine data of various types when deriving source weights and truths. Under this general framework, we discussed several common data types and loss functions, derived effective solutions, and analyzed its convergence. Different from existing truth discovery approaches that focus on facts [9, 12, 18, 22] or continuous data [23], the proposed CRH model learns source reliability degrees jointly from various properties with different data types. Unique characteristics of each data type is considered, and all types contribute to source reliability estimation together. This joint inference improves source reliability estimation and leads to better truth discovery on heterogeneous data.

3. EXPERIMENTS

In this section, we report the experimental results on both real-world and simulated data sets, which show that the proposed CRH method is efficient and outperforms state-of-the-art conflict resolution methods when integrating multiple sources of heterogeneous data. We first discuss the experiment setup in Section 3.1, and then present experimental results in Section 3.2.

3.1 Experiment Setup

In this part, we present the performance measures and discuss the baseline methods.

3.1.1 Performance Measures

The problem setting is that we have multi-source input and the ground truths. All the conflict resolution methods are conducted in an unsupervised manner in the sense that the ground truths will only be used in evaluation. In this experiment, we focus on two types of data: categorical and continuous. To evaluate the performance of various conflict resolution methods, we adopt the following measures for these two data types:

- Error Rate: For categorical data, we use *Error Rate* as the performance measure of an approach, which is computed as the percentage of the approach's output that are different from the ground truths.
- MNAD: For continuous data, we can measure the overall absolute distance from each method's output to the ground truths, which indicates how close the output are to the ground truths. As different entries may have different scales, we normalize the distance on each entry by its own variance, and then calculate their mean. This leads to the measure *Mean Normalized Absolute Distance* (*MNAD*).

For both measures, the **lower** the value, the closer the method's estimation is to the ground truths and thus the **better** the performance.

3.1.2 Baseline Methods

For the proposed CRH method, we use weighted voting (Eq(13)) for categorical data due to its time and space efficiency. On continuous data, we use weighted median (Eq(23)), which is efficient and robust in noisy environment with outliers. Weight assignment is computed by the inverse logarithm of the ratio between the deviation to the truth and the maximum distance so that the difference in source reliability is emphasized. We compare the proposed approach with the following baseline methods that cover a wide variety of ways to resolve conflicts. These approaches can be partitioned into three categories.

Conflict Resolution Methods Applied on Continuous Data Only. The following approaches can only be applied on continuous data, and thus they will ignore the input from categorical properties.

- Mean: Mean simply takes the mean of all observations on each property of each object as the final output.
- Median: Median calculates the median of all observations on each property of each object as the final output.

- GTM [23]: Gaussian Truth Model (GTM) is a Bayesian probabilistic approach especially designed for solving conflict resolution problem on continuous data. Note that this approach only uses partial data (continuous) while other truth discovery methods use all the data (categorical and continuous), and thus insufficient data may lead to performance degrade in GTM compared with others. However, we still include GTM in the comparison as it is an important truth discovery method applied on continuous data.

Among them, Mean and Median are traditional conflict resolution approaches, and GTM is a truth discovery approach which considers source reliability estimation.

Conflict Resolution Methods Applied on Categorical Data Only. We apply majority voting approach on categorical properties only. This is the traditional way of resolving conflicts in categorical data without source reliability estimation.

- Voting: The value which has the highest number of occurrences is regarded as the fused output.

Conflict Resolution Methods by Truth Discovery. Many of the existing conflict resolution approaches that consider source reliability (often referred to as "truth discovery" approaches) are developed to find true "facts" for categorical properties. However, we can enforce them to handle data of heterogeneous types by regarding continuous observations as "facts" too.

- Investment [18]: In this approach, a source "invests" its reliability uniformly on the observations it provides, and collects credits back from the confidence of those observations. In turn, the confidence of an observation grows according to a non-linear function defined based on the sum of invested reliability from its providers.
- PooledInvestment [18]: PooledInvestment is similar to Investment. The only difference is that the confidence of an observation is linearly scaled instead of non-linearly scaled.
- 2-Estimates [12]: This approach is proposed based on the assumption that "there is one and only one true value for each entry". If a source provides an observation for an entry, 2-Estimates assumes that this source votes against different observations on this entry.
- 3-Estimates [12]: 3-Estimates improves 2-Estimates by considering the difficulty of getting the truth for each entry, the estimation of which will affect the source's weight.
- TruthFinder [22]: TruthFinder adopts Bayesian analysis, in which for each observation, its confidence is calculated as the product of its providers' reliability degrees. Similarity function is used to adjust the vote of a value by considering the influences between facts.
- AccuSim [9]: AccuSim also applies Bayesian analysis and it also adopts the usage of the similarity function. Meanwhile, it considers complement vote which is adopted by 2-Estimates and 3-Estimates. Note that in [9], other algorithms have been proposed to tackle source dependency issues in resolving conflicts, which are not compared here because we do not consider source dependency in this paper but leave it for future work.

The comparison between the proposed framework with these baseline approaches on heterogeneous data can show that 1) using both types of data jointly gives better source reliability estimation than using individual data types separately, but 2) an accurate weight can only be obtained by taking unique characteristics of each data type into consideration.

We implement all the baselines and set the parameters according to their authors' suggestions. All the experimental results in this section except for MapReduce experiments are conducted on a Windows machine with 8G RAM, Intel Core i7 processor.

3.2 Experimental Results

In this section, by comparing the proposed CRH approach with the baseline methods, we show the power of simultaneously modeling various data types in a joint framework on both real-world and simulated data sets. We also show the efficiency of the proposed approach on single machine and Hadoop cluster.

3.2.1 Real-world Data Sets

We use three real-world data sets to demonstrate the effectiveness of the proposed method.

Weather Forecast Data Set. Weather forecast integration task is a good test bed because the data contains heterogeneous types of properties. Specifically, we integrate weather forecasting data collected from three platforms: Wunderground[3], HAM weather[4], and World Weather Online[5]. On each of them, we crawl the forecasts of three different days as three different sources, so altogether there are nine sources. For each source, we collected data of three properties: high temperature, low temperature and weather condition, among which the first two are continuous and the last is categorical. To get ground truths, we crawl the true weather information for each day. We collected the data for twenty US cities over a month.

Stock Data Set. The stock data [15], crawled on every work day in July 2011, consists of 1000 stock symbols and 16 properties from 55 sources, and the ground truths are also provided. Here, we treat the data set as heterogeneous. More specifically, property *volume*, *Shares outstanding* and *Market cap* are considered as continuous type, and the rest ones are considered as categorical type.

Flight Data Set. The flight data [15], crawled over one-month period starting from December 2011, consists of 1200 flights and 6 properties from 38 sources. We conduct pre-processing on the data to convert the gate information into the same format and the time information into minutes. The ground truths are also available. In this work, we show results on the flight data by treating gate information as categorical type and time information as continuous type. Note that we have a different task setting compared with [15] for Stock and Flight data when we treat them as heterogeneous types.

Table 5 shows the statistics of these three data sets. Note that the number of entries does not equal to the number of ground truths because we only have a subset of entries labeled with ground truths. The ground truths are not used by any of the approaches, but only used in the evaluation.

Table 5: Statistics of Real-world Data Sets

	Weather Data	Stock Data	Flight Data
# Observations	16038	11748734	2790734
# Entries	1920	326423	204422
# Ground Truths	1740	29198	16572

In Table 6, we summarize the performance of all the methods in terms of *Error Rate* on categorical data and *MNAD* on continuous data for three real-wold data sets. Although our approach outputs truths on both data types simultaneously, we evaluate the

[3] http://www.wunderground.com

[4] http://www.hamweather.com

[5] http://www.worldweatheronline.com

Table 6: Performance Comparison on Real-world Data Sets

	Weather Data		Stock Data		Flight Data	
Method	Error Rate	MNAD	Error Rate	MNAD	Error Rate	MNAD
CRH	**0.3759**	**4.6947**	**0.0700**	**2.6445**	**0.0823**	**4.8613**
Mean	NA	4.7840	NA	7.1858	NA	8.2894
Median	NA	4.9878	NA	3.9334	NA	7.8471
GTM	NA	4.7914	NA	3.4253	NA	7.6703
Voting	0.4844	NA	0.0817	NA	0.0859	NA
Investment	0.4913	5.2361	0.0983	2.8081	0.0919	6.4153
PooledInvestment	0.4948	5.5788	0.0990	2.7940	0.0925	5.8562
2-Estimates	0.5327	5.5258	0.0726	2.8509	0.0885	7.4347
3-Estimates	0.4810	5.1943	0.0818	2.7749	0.0881	7.1983
TruthFinder	0.4586	5.1293	0.1194	2.7140	0.0950	8.1351
AccuSim	0.4672	5.0862	0.0726	2.8503	0.0881	7.3204

performance separately on these two data types due to the different measures for different data types. It can be seen that the proposed CRH approach achieves better performance on both types of data compared with all the baselines. For example, on weather data, the number of mismatches from ground truths drops from 266 (the best baseline) to 218 out of 580 entries by using CRH (on categorical data). On Stock and Flight data sets where baselines have already achieved good performance, we still can see the performance improvement of CRH over the best baseline (1719 ->1657 out of 23677, and 427 ->414 out of 4971). Similarly, the gain on continuous data can be consistently observed on all three data sets.

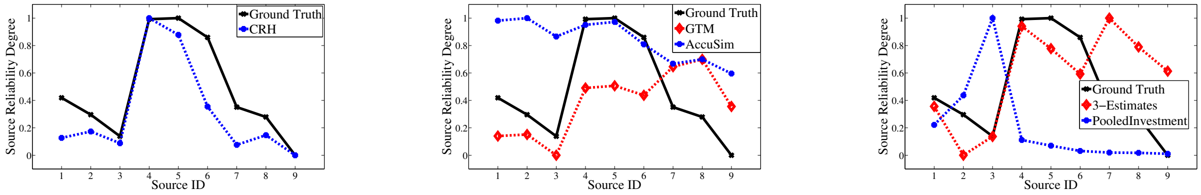
By outperforming various conflict resolution approaches applied separately on categorical data and continuous data, the proposed CRH approach demonstrates its advantage in modeling source reliability more accurately by jointly inferring from both types of data. GTM can not estimate source reliability accurately merely by continuous data which may not have sufficient information. This also justifies our assumption that each source's reliability on continuous and categorical data is consistent so the estimation over different data types complements each other.

The reason that the proposed CRH approach beats the other conflict resolution approaches that are applied on both types of data is that these approaches cannot capture the unique characteristics of each data type. This is further supported by the fact that the performance of those approaches is relatively better on categorical data, but deviates more from the truths on the continuous data. In contrast to existing approaches, the proposed CRH framework can take data type into consideration, which will provide a better estimation of source reliability, and thus result in more accurate truth estimation.

As source reliability is the key to obtain correct truths, we further show the source reliability degrees estimated for the 9 sources by various approaches on the weather forecast data set. We choose this data set because it consists of nine sources only, which is more practical to demonstrate. We first compute the true source reliability by comparing the observations made by each source with the ground truths. Reliability of a source is defined as the probability that the source makes correct statements on categorical data, and the chance that the source makes statements close to the truth on continuous data. To simplify the presentation, we combine the reliability scores of continuous and categorical data into one score for each source. To make it clear, we show the source reliability degrees in three plots presented in Figure 1, each of which shows the comparison between the ground truths and some of the approaches.

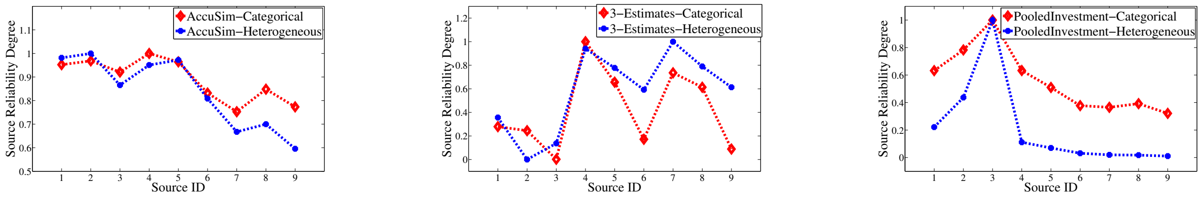
Figure 1(a) shows that the source reliability degree estimated by CRH method is in general consistent with that obtained from the ground truths. By effectively characterizing different data types in a joint model, the proposed approach can successfully distinguish good sources from bad ones, and accordingly derive the truth based on good sources. In Figures 1(b) and 1(c), we show the reliability degrees of 9 sources estimated by GTM, AccuSim, 3-Estimates and PooledInvestment compared with the ground truth reliability. We particularly show the results on these approaches because 3-Estimates and PooledInvestment are improved solutions compared with 2-Estimates and Investment respectively claimed in the corresponding papers, and TruthFinder has similar performance with AccuSim on this data set. As different methods adopt various functions to estimate the source reliability scores, to make them comparable, we normalize all the scores into the range $[0, 1]$. Among these approaches, 3-Estimates and GTM calculate the unreliability degrees, so we convert their scores to reliability degrees to show the comparison. The plots show that the baseline methods can capture the difference among sources in making accurate claims to a certain extent, but the patterns in source reliability detected by them are not very consistent with the ground truths, which can thus explain the increased error in truth detection in Table 6.

Note that the improved performance of the proposed approach is achieved when each individual data type's characteristics are fully taken into account in the joint modeling of heterogeneous data. To demonstrate this point, we show that using heterogeneous data without distinguishing data types cannot help improve the performance. As shown in Table 7, the baseline approaches, which regard all data types the same, have similar performance when they are applied on heterogeneous data and on categorical data only. This indicates that the incorporation of continuous data does not make difference for these approaches. To further investigate this behavior, we plot the source weights obtained by AccuSim, 3-Estimates, and PooledInvestment on categorical data and heterogeneous data respectively in Figure 2. From the plots, we can see that for any of these approaches, these two groups of source reliability scores are close to each other. The reason behind this phenomenon is that unlike categorical data, continuous data are noisier and wider distributed, and sources disagree much more on the values. When these approaches ignore the characteristics of different data types, they cannot find meaningful truths. Thus, the deviations are rather random and uniform across different sources. When summing up the deviations on both categorical and continuous data to compute source weights, the inputs from the continuous side do not significantly affect the source weight distribution because it is quite similar across sources. In other words, the incorporation of continuous data without considering their characteristics can hardly benefit or even harm the weight estimation. Therefore, these approaches get similar source weight distributions with and without continuous data and the results are not improved. This phenomenon further proves that only utilizing heterogeneous data by considering the characteristics of different data types in source reliability estimation can obtain a good result. That is the reason why the proposed approach performs better than baseline approaches.



(a) CRH (b) Baselines (c) Baselines

Figure 1: Comparison of Source Reliability Degrees with Ground Truths



(a) AccuSim (b) 3-Estimates (c) PooledInvestment

Figure 2: Comparison of Source Reliability Degrees between Methods Applied on Categorical and Heterogeneous Data

Table 7: Comparison for Baselines on Weather Data

	Heterogeneous Data		Categorical Data	
Method	Error Rate	MNAD	Error Rate	MNAD
Investment	0.4913	5.2361	0.4776	NA
PooledInvestment	0.4948	5.5788	0.4914	NA
2-Estimates	0.5327	5.5258	0.5121	NA
3-Estimates	0.4810	5.1943	0.4638	NA
TruthFinder	0.4586	5.1293	0.4603	NA
AccuSim	0.4672	5.0862	0.4620	NA

3.2.2 *Noisy Multi-source Simulations*

To further demonstrate the advantages of the proposed framework in the environment involving various reliability degrees and different loss functions, we conduct experiments on simulated data sets generated from UCI machine learning data sets. We choose two data sets: UCI Adult[6] and Bank[7] data sets, each of which has both continuous and categorical properties. The original data sets are regarded as the ground truths, and based on each of them, we generate a data set consisting of multiple conflicting sources by injecting different levels of noise into the ground truths as the input to our approach and baseline methods. Gaussian noise is added to each continuous property, and values in categorical properties are randomly flipped to generate facts that deviate from the ground truths. To better simulate the real-world data, we round the continuous type data based on their physical meaning. A parameter α is used to control the reliability degree of each source (a lower α indicates a lower chance that the ground truths are altered to generate observations). In this way, we generate data sets which contain 8 sources with various degrees of reliability ($\alpha = \{0.1, 0.4, 0.7, 1, 1.3, 1.6, 1.9, 2\}$). Table 8 shows the statistics of these two data sets.

Table 8: Statistics of Simulated Data Sets

	Adult Data	Bank Data
# Observations	3646832	5787008
# Entries	455854	723376
# Ground Truths	455854	723376

Tables 9 summarizes the results of all the approaches on these two data sets. It can be seen that CRH can fully recover all the truths on categorical data, and find the true value for continuous data with very small distance by inferring accurate source reliability degrees. Similar to the experiments on the weather data set, we can still observe the great improvement in truth detection performance compared with baseline approaches due to the proposed method's advantage in source reliability estimation. Existing approaches cannot provide accurate estimate of source reliability because they either take incomplete data (only categorical or continuous), or do not model the characteristics of both data types jointly.

Table 9: Performance Comparison on Simulated Data Set

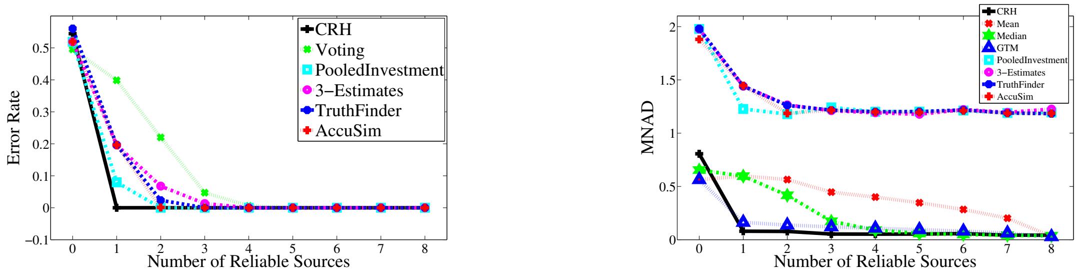
	Adult Data		Bank Data	
Method	Error Rate	MNAD	Error Rate	MNAD
CRH	**0.0000**	**0.0637**	**0.0000**	**0.0789**
Mean	NA	0.3673	NA	0.3671
Median	NA	0.2470	NA	0.2491
GTM	NA	0.0810	NA	0.0948
Voting	0.1029	NA	0.2314	NA
Investment	0.0530	0.1391	0.1197	0.1588
PooledInvestment	0.0215	0.1008	0.0241	0.0866
2-Estimates	0.0497	0.1355	0.1152	0.1583
3-Estimates	0.0497	0.1355	0.1152	0.1583
TruthFinder	0.0346	0.1272	0.1097	0.1589
AccuSim	0.0288	0.1145	0.0681	0.1571

On these simulated data sets, we also investigate how the performance of the proposed CRH approach varies with respect to different distributions of source reliability degrees. To illustrate the effect more clearly, we choose two reliability degrees: $\alpha = 0.1$ and $\alpha = 2$, which correspond to reliable and unreliable sources respectively. We now fix the total number of sources as 8, and change the number of reliable sources to conduct a series of experiments. Figures 3 and 4 show the performance of the proposed approach and baseline methods on Adult and Bank data sets respectively. In each of them, we show the performance on categorical and continuous data respectively when we vary the number of reliable sources from 0 to 8 (out of 8 sources in total).

The following observations can be made from the results: 1) The plots support our previous findings that the CRH framework outperforms existing conflict resolution techniques, which ignore the unique characteristics of each data type. When sources are equally reliable or unreliable (number of reliable sources equals to 0 or 8), the CRH model achieves similar performance as that of vot-

[6] http://archive.ics.uci.edu/ml/datasets/Adult

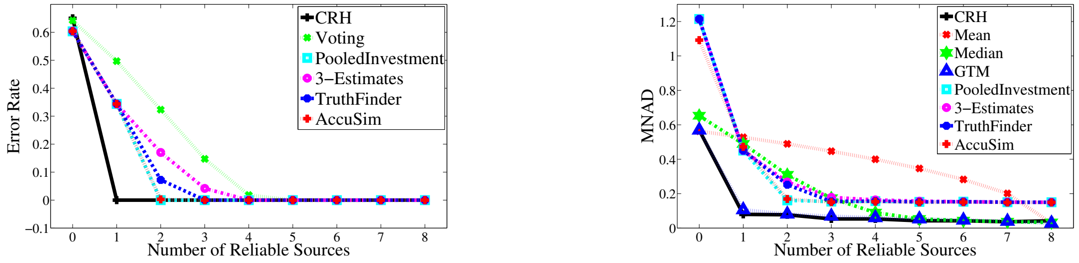
[7] http://archive.ics.uci.edu/ml/datasets/Bank+Marketing



(a) Categorical Data

(b) Continuous Data

Figure 3: Performance w.r.t. # Reliable Sources on Adult Data Set



(a) Categorical Data

(b) Continuous Data

Figure 4: Performance w.r.t. # Reliable Sources on Bank Data Set

Table 10: Comparison for Different Loss Functions

	Method	Error Rate	MNAD
Scenario 1	Eq(13) + Eq(20)	0.0125	0.1533
	Eq(13) + Eq(23)	0.0125	0.1938
	Eq(17) + Eq(20)	0.0375	0.1628
	Eq(17) + Eq(23)	0.0125	0.1938
Scenario 2	Eq(13) + Eq(20)	0.0375	0.3030
	Eq(13) + Eq(23)	0.0250	0.2052
	Eq(17) + Eq(20)	0.0375	0.2899
	Eq(17) + Eq(23)	0.0250	0.2052

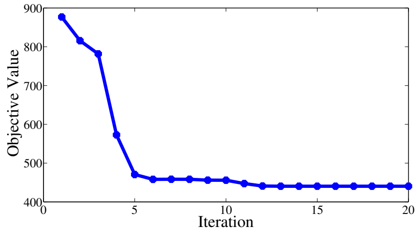


Figure 5: Convergence Rate

ing/averaging approaches. However, when the reliability degree varies across sources, CRH performs much better. 2) In general, it is easier to detect truths when we have a bigger number of reliable sources. However, on categorical data, even when only 1 out of 8 sources is reliable, CRH can still discover most of the truths. Clearly, the proposed approach can successfully infer source reliability and thus detect the truths that are stated by the minority. 3) On continuous data, we can see that the convergence rate is slower than that on categorical data. Conflict resolution on continuous data is in general more difficult due to the higher complexity of the truth space and more complicated definition of closeness to the truths.

Next we discuss the effect of loss functions and give some guidelines on how to choose appropriate loss functions under different scenarios (Table 10).

Under Scenario 1, we generate five sources with different levels of noise and the levels of noise are relatively low on all sources ($\alpha = \{0.5, 0.5, 0.5, 1, 1\}$), which means that sources have low variance on continuous data and tend to agree with each other on categorical data. The combination of Eq(13) and Eq(20) achieves the best performance (using traditional 0-1 and squared loss). Under Scenario 2, we add more noise on two out of the five sources ($\alpha = \{0.5, 0.5, 1, 3.5, 4.5\}$). These two sources contain outliers on continuous data, and many mistakes on categorical data. We find that Eq(13) + Eq(23) and Eq(17) + Eq(23) work better than the others. This is because that normalized absolute deviation gives a smaller penalty on outliers than normalized squared loss does, and thus is more tolerant of outliers.

3.2.3 *Efficiency*

In this section, we study the efficiency of the proposed CRH framework. We first explore its convergence rate in practice, and then show its running time on both single machine and Hadoop cluster.

Convergence Rate. As we proved in Theorem 6, the convergence of CRH framework is guaranteed when Eq(4) is used as constraint, Eq(16) or/and Eq(19) is/are used as loss functions. We further demonstrate the convergence rate using weather data set. Figure 5 shows the change of the objective value with respect to each iteration. We can see that the objective value decreases fast at the first five iterations and then reaches a stable stage, showing that the proposed method converges quickly in practice.

Running Time. We sample different number of observations and entries from weather data set to show the running time of the CRH framework on single machine. As shown in Table 11, the proposed approach has linear complexity in the number of observations. To further demonstrate this, we compute Pearson's correlation coefficient, which is a commonly used method to test linear relationship between variables. It ranges from 1 to -1. The closer it is to 1 (or -1), the stronger positive (or negative) linear relationship the variables have. In our experiment, the Pearson's correlation coefficient for running time and the number of observations is higher than 0.99, which indicates that they are highly linearly correlated.

As mentioned in the discussions of the CRH framework, it can be easily adapted to MapReduce programming model. As the number of observations can be easily varied in a large range, we use the simulated data sets to evaluate the running time of CRH on Hadoop cluster. Based on the Adult data set, we generate large-scale data sets by adding different noise levels on the original data set as we discussed. The number of observations vary from 10^4 to 10^8. The proposed CRH framework is implemented using MapReduce model. The experiments are conducted on a 4-node Dell Hadoop cluster with Intel Xeon E5-2403 processor (4x 1.80 GHz, 48 GB RAM). As shown in Table 11, the fusion process using the MapReduce version of the proposed approach can finish in a short time. The running time mainly comes from the setup overhead when the number of observations is not very large, but the speed-up in the execution time is more obvious when the number of observations increases. For example, on a data set with size 10^8, the whole process only took 669s.

Table 11: Running Time

Single Machine		Hadoop Cluster	
# Observations	Time (s)	# Observations	Time (s)
5×10^4	1.5575	1×10^4	94
7×10^4	2.3265	1×10^5	96
9×10^4	2.9505	1×10^6	100
1.2×10^5	3.4133	1×10^7	193
1.4×10^5	4.3827	1×10^8	669
1.6×10^5	4.6724	4×10^8	1384
Pearson Correlation	0.9903	Pearson Correlation	0.9811

4. CONCLUSIONS

To extract insightful knowledge from an overwhelming amount of information generated by numerous industries, it is crucial to automatically identify trustworthy information and sources from multiple conflicting data sources. As heterogeneous data is ubiquitous, a joint estimation on various data types can lead to better estimation of truths and source reliability. However, existing conflict resolution work either regards all the sources equally reliable, or models different data types individually. Therefore, we propose to model the conflict resolution problem on data of heterogeneous types using a general optimization framework called CRH that integrates the truth finding process on various data types seamlessly. In this model, truth is defined as the value that incurs the smallest weighted deviation from multi-source input in which weights represent source reliability degrees. We derive a two-step iterative procedure including the computation of truths and source weights as a solution to the optimization problem. The advantage of this framework is its ability of taking various loss and regularization functions to characterize different data types and weight distributions effectively. We derive efficient computation approach that is linear with respect to the number of observations and the approach can be easily implemented in MapReduce model. We conduct experiments on weather, stock and flight data sets collected from multiple platforms as well as simulated multi-source data generated from UCI machine learning data sets. Results demonstrate the efficiency and the advantage of the proposed CRH approach over existing conflict resolution approaches in finding truths from heterogeneous data. In the future, we plan to adapt the framework to more complicated conflict resolution scenarios, such as the cases involving multiple truths and source dependency.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *JMLR*, 6:1705–1749, 2005.

[2] D. P. Bertsekas. *Non-linear programming*. Athena Scientific, 1999.

[3] L. Blanco, V. Crescenzi, P. Merialdo, and P. Papotti. Probabilistic models to reconcile complex data from inaccurate data sources. In *Proc. of CAiSE*, pages 83–97, 2010.

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

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

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

[7] C.-T. Chu, S. K. Kim, Y.-A. Lin, Y. Yu, G. Bradski, A. Y. Ng, and K. Olukotun. Map-reduce for machine learning on multicore. In *NIPS*, pages 281–288, 2006.

[8] C. Dai, D. Lin, E. Bertino, and M. Kantarcioglu. An approach to evaluate data trustworthiness based on data provenance. In *Proc. of SDM*, pages 82–98, 2008.

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

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

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

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

[13] Z. Jiang. A decision-theoretic framework for numerical attribute value reconciliation. *TKDE*, 24(7):1153–1169, 2012.

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

[15] X. Li, X. L. Dong, K. B. Lyons, W. Meng, and D. Srivastava. Truth finding on the deep web: Is the problem solved? *PVLDB*, 2013.

[16] A. Marian and M. Wu. Corroborating information from web sources. *IEEE Data Engineering Bulletin*, 34(3):11–17, 2011.

[17] J. Nocedal and S. Wright. *Numerical optimization*. Springer, 2006.

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

[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] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *JOTA*, 109(3):475–494, 2001.

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

[22] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. In *Proc. of KDD*, pages 1048–1052, 2007.

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

[24] 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.