A Probabilistic Model for Estimating Real-valued Truth from Conflicting Sources

Bo Zhao Jiawei Han
Department of Computer Science, University of Illinois, Urbana, IL, USA
{bozhao3, hanj}@illinois.edu

ABSTRACT

One important task in data integration is to identify truth from noisy and conflicting data records collected from multiple sources, *i.e.,* the *truth finding problem*. Previously, several methods have been proposed to solve this problem by simultaneously learning the quality of sources and the truth. However, all those methods are mainly designed for handling categorical data but not numerical data. While in practice, numerical data is not only ubiquitous but also of high value, e.g. price, weather, census, polls, economic statistics, *etc.* Quality issues on numerical data can also be even more common and severe than categorical data due to its characteristics. Therefore, in this work we propose a new truth-finding method specially designed for handling numerical data. Based on Bayesian probabilistic models, our method can leverage the characteristics of numerical data in a principled way, when modeling the dependencies among source quality, truth, and claimed values. Experiments on two real world datasets show that our new method outperforms existing state-of-the-art approaches.

1. INTRODUCTION

In the era of big data there are almost always multiple sources providing data or claims with regard to the same entities, ranging from presidential election polls, census, and economic statistics to stock price predictions and weather forecasts. However, the cost that comes along with such data abundance is that data is often not consistent across multiple sources and people do not know which sources are more trustworthy and what is the truth, so it would be extremely helpful if an algorithm can automatically learn the quality of sources and derive the truth from the data. Several proposals have been made in the past [4, 6, 10, 11, 13–15] to solve this *truth-finding* problem. However, it is worth mentioning that all the previous methods are mainly developed to handle categorical input, such as authors of books or cast members of movies, but when applied on numerical data, those methods do not consider much its unique characteristics and therefore the performance is likely to be not optimal. While in practice, numerical data is as ubiquitous and important as nominal data, if not more; and the data quality issue could be even more common and severe, which is the motivation of this work: a new truth finding method specially designed for handling numerical data.

Numerical data has several properties that need to be treated carefully by the truth-finding method. First, different from the standard categorical setting where different claimed values do not have much correlation, distance or similarity can be defined between numerical claims of the same entities and it should be considered during the inference of truth. For example, if three claims with values of 100, 90 and 50 are observed, it is more likely that the truth is closer to 100 and 90 rather than 50. If we assume most sources are benign and tend to tell the truth in general, then it is reasonable to give claims that are closer to the truth higher probability to be observed than those that are farther. Then, finding truth that can maximize the likelihood of observed data becomes equivalent to searching for the truth that minimizes its certain form of overall distance to all the claimed values.

Second, it is not sufficient to simply define quality of a source as how often its claims are exactly right or wrong, because intuitively claims that are closer to the truth should get more credit for the source than those that are farther. On the other hand, assuming the source quality is known, if two claims have the same distance to the truth, the one made by the source with lower quality should have higher probability of being observed than the other one, since lower quality sources tend to give claims deviating more from the truth than higher quality sources. Such probabilistic modeling can lead to a desired weighing scheme during the truth inference: claims made by high quality sources will have higher weights in deciding where the truth should be than unreliable sources.

Third, the consensus level among claims for each entity should be a factor in estimating truth and source quality. Specifically, for entities that sources generally do not agree on, the inferred truth should get lower confidence and sources making different claims should be punished less compared with cases where high consensus can be achieved among most sources. Furthermore, numerical values for different entities can be in different scales so proper normalization needs to be deployed to prevent biased estimation of source quality. For example, population claims about big cities are more likely to deviate more from the truth in terms of the absolute difference, which may cause sources that contain more big cities get unfair punishment if the data scale is not properly normalized.

Last but not least, in numerical data outliers can happen more often and cause more severe damages to model assumptions of the truth-finding methods if they are not effectively detected. For instance, the mean of observed data can shift infinitely from the truth due to only one outlier. Therefore, when we design a truth-finding method that adapts to the characteristics of numerical data, it is also important to detect outliers and reduce their damages as much as possible. In most cases, it is possible to make an initial estimate

Table 1: A sample census database.

Entity (City)	*Value (Population)*	*Source*
New York City	8,346,794	Freebase
New York City	8,244,910	Wikipedia
New York City	8,175,133 (truth)	US Census
New York City	7,864,215	BadSource.com
Urbana	36,395 (truth)	US Census
Urbana	36,395 (truth)	Wikipedia
Urbana	34,774	Freebase
Urbana	1,215	BadSource.com
Los Angeles	2,966,850	Freebase
Los Angeles	3,364,215	BadSource.com
...	...	...

of the truth and leverage it as prior for detecting outliers.

EXAMPLE 1. *Table 1 shows a sample census database with city population claims; truth is also labeled although in practice it is unknown. It is clear in this example that Freebase should have better quality than BadSource.com since its claims are closer to the truth, although neither of them make exactly correct claims. As a result, when inferring the true population of Los Angeles, Freebase should have higher weights. Moreover, Freebase's claim on New York City should not incur more punishment than Urbana simply because its absolute difference from the truth is larger, ignoring the fact that New York city is a much bigger city; and in this case, the opposite seems more reasonable since the general consensus on Urbana's population is higher. Also notice that claiming the population of Urbana is 1,215 is an outlier, and it should be easily detected by algorithms. However, if it is not detected and still treated as a reasonable claim, the truth, which is the highest value among all claims about the city, will be unnecessarily assigned lower confidence and the credit of US Census and Wikipedia will also be harmed.*

To address the issues of truth finding on numerical data we just discussed, in this work we propose a Bayesian probabilistic model we call the *Gaussian Truth Model (*GTM*)*, which can leverage the characteristics of numerical data in a principled manner, and infer the real-valued truth and source quality without any supervision. To the best of our knowledge, this is the first truth-finding method designed for numerical data.

The Bayesian nature of our model makes it possible to leverage the output of any other truth finding methods as prior belief of the truth, which could be beneficial since some non-numerical methods may also be less affected by outliers, thus their output can provide a better initial guess in our model to further reduce the damage of outliers.

Bayesian priors on source quality can also be incorporated for smoothing quality estimates for sources with very few claims, since the maximum likelihood estimates on small volume data is often inaccurate. The additional benefit of having source quality priors is that it allows our method to run incrementally if data comes in a stream.

Experiments on two real world datasets (Wikipedia edit history of city population and people biographies) demonstrate GTM outperforms state-of-the-art methods.

In the following sections, we first describe our data model and formalize the problem in Section 2. We then introduce the Gaussian truth model in Section 3. Section 4 presents our experimental results. Related work are discussed in Section 5. Finally, we conclude the paper in Section 6.

2. PROBLEM FORMULATION

2.1 Data Model

Definition 1. Let $\mathcal{C} = \{c_1, c_2, ..., c_C\}$ be the set of *claims* GTM takes as input. Each claim c is in the format of (e_c, s_c, v_c), where e_c is the *entity* of the claim, s_c is the *source* of the claim, and v_c is the numerical *value* of the claim. Let $\mathcal{S} = \{s_1, s_2, ..., s_S\}$ be the set of sources that appear in $\mathcal{C}$, and let $\mathcal{E} = \{e_1, e_2, ..., e_E\}$ be the set of entities that appear in $\mathcal{C}$, and let $\mathcal{C}_e$ be the set of claims that are associated with entity e.

Table 1 is an example of a set of input claims.

Definition 2. Let $\mathcal{T} = \{(e_1, t_1), (e_2, t_2), ..., (e_E, t_E)\}$ be a set of *truths*, where each truth t is a real value associated with one entity in $\mathcal{E}$. For each $e \in \mathcal{E}$, we denote the truth associated with e as t_e.

Truth set $\mathcal{T}$ is the output format of the truth finding methods. A human generated truth set is used for evaluating the effectiveness of different approaches.

2.2 Problem Definitions

We can now define the problems of interest in this paper.

Inferring truth. Given an input claim set $\mathcal{C}$ with no truth information, we want to output the inferred truths $\mathcal{T}$ for all entities $\mathcal{E}$ contained in $\mathcal{C}$.

Inferring source quality. Besides the true values for each entity, we also would like to automatically infer quality information for each source represented in $\mathcal{C}$. Such source quality information represents how close each source's claims are to the truth in general; it not only can help people understand the data sources or uncover potential problems during the data collection phase, but also can be used as prior parameters for inferring truth from new data to prevent rerunning the model on the whole dataset, *etc.*

The truth values and source quality are closely related and are estimated simultaneously by GTM. More details will be explained in the next section.

3. GAUSSIAN TRUTH MODEL

In this section we will formally introduce our proposed model, called the Gaussian Truth Model, for estimating the real-valued truth and the quality of data sources. We will first discuss the intuitions behind our model, then provide details about the major components of the approach including how the probabilistic model is constructed as well as the outlier detection and data normalization step, and how our intuitions are embedded in the model.

3.1 Intuitions

3.1.1 Real-valued Truths and Claims

In the numerical setting, distance between real values should be considered since it provides richer information about where the truth should be. For example, if three claims with values of 100, 90 and 50 are observed, it is more likely that the truth is closer to 100 and 90 rather than 50, but such conclusion can not be made if the values are treated as nominal ones and their distance is not considered.

To leverage the distance in a principled probabilistic way, we model the truth of each entity as an unknown random variable that takes real values, and use it as the mean parameter in the probabilistic distribution that models the probability of observing each claimed value of the entity. Such probabilistic distribution should

satisfy the property we have discussed earlier that values closer to the mean parameter, *i.e.*, the truth, have higher probability than values that are farther. Essentially many distributions have this property, and in GTM we choose the Gaussian distribution since it is the most commonly used distribution to model the generation of real-valued errors due to its quadratic penalty in logarithm form which leads to more efficient inference algorithms. With such modeling, truth values that maximize the data likelihood are essentially values with minimum squared deviation to each claimed value.

Observed data can often be in small volume or contain outliers that can introduce bias in the maximum likelihood estimation (MLE). Therefore, Bayesian methods often leverage prior distributions of model parameters to obtain maximum a posteriori (MAP) estimates. In GTM, any initial estimation of the truth such as the mean, the median, the most frequent value, or output of any other truth-finding algorithms can be incorporated as priors to the model. Some of these priors are less affected by outliers compared with the average, and therefore can help improve the performance of GTM in practice. It is also worth mentioning that the prior's weight needs to be automatically adjusted based on the data scale and the consensus level of each entity. Specifically, if the claimed values generally tend to not agree with the prior, then the truth should get more freedom for deviating from the prior belief.

3.1.2 Quality of Sources

Since distance can be defined between real-valued claims and the truth, the quality of sources should naturally correspond to how close their claims are to the truth in general rather than how often their claims are exactly correct. We have mentioned that the probability of observing each claim is given by a Gaussian distribution with the truth as its mean parameter; and the variance parameter of this Gaussian distribution actually controls how likely claims deviate from the mean, which exactly relates to the quality of the claim's source. Intuitively, inaccurate sources are more likely to make claims that deviate more from the truth, and therefore they corresponds to larger variance; high quality sources correspond to lower variance, respectively.

Formally speaking, in GTM we model the quality of each source as a random variable that takes real values, and for each claim we use the quality of its source as the variance parameter in the Gaussian distribution that assigns probability to the claimed value.

With such modeling, the quality of sources can be easily estimated from the data if the truth is known. On the other hand, if the source quality is known, truth can also be inferred and inaccurate sources will be given lower weights in deciding where the truth should be, because the cost of deviating from claims made by low quality sources is discounted by their high variance in the Gaussian function. We can see such iterative computation of truth and source quality is similar to many previous unsupervised truth-finding methods, but the difference is that in GTM the characteristics of numerical data can be well leveraged in a principled way.

There are a few additional issues in modeling source quality. One is that the data scale and consensus level of each entity should be considered. Intuitively, if for one entity claims from all sources have higher deviation from the truth in general, either because the truth is a larger value or it is more controversial, each source should get less punishment for the same amount of error. Another point worth mentioning is that in GTM source quality is also modeled in the Bayesian fashion, and therefore it is easy to incorporate any prior knowledge about the quality of all sources or specific ones either provided by domain experts or output by GTM on historical data. Even such knowledge is not available, specifying prior has the effect of smoothing and helps prevent overfitting.

Algorithm 1 Data Preprocessing

```
{Outlier Detection}
for all e ∈ E do
    {Based on relative errors and absolute errors.}
    for all c ∈ C_e do
        if |v_c − t̂_e|/t̂_e > δ_0 or |v_c − t̂_e| > δ_1 then
            outlier[c] ← True
    {Based on z-scores (Gaussian p-values).}
    σ̂_e ← standard_deviation(C_e, outlier)
    repeat
        new_outlier ← False
        for all c ∈ C_e do
            if |v_c − t̂_e|/σ̂_e > δ_2 then
                outlier[c] ← True
                new_outlier ← True
        σ̂_e ← standard_deviation(C_e, outlier)
    until new_outlier = False
{Normalization: calculating z-scores.}
for all e ∈ E do
    for all c ∈ C_e do
        o_c ← (v_c − t̂_e)/σ̂_e
```

3.2 Preprocessing

Before introducing details of the model, we first explain the preprocessing step including how the input claims are normalized to prevent biased estimation, and how outliers are detected.

In the previous section, we have mentioned that there are various estimates of the truth that can be leveraged by GTM as priors, such as the mean, the median, the most frequent value or output of any other truth-finding methods. In fact, the prior information is also utilized for normalizing the data and detecting outliers during the preprocessing step.

Based on robust statistics, the sample mean can be shifted infinitely by outliers and therefore is not a good prior, but some measures, such as the median or the output of non-numerical truth-finding algorithms, would be more robust and suitable to serve as the prior. With the prior given, any claimed values deviating too far from the prior can be treated as outliers, and there are various ways to measure the deviation, such as relative errors, absolute errors, Gaussian p-values, *etc.*

Let $\hat{\mathcal{T}} = \{(e_1, \hat{t}_1), (e_2, \hat{t}_2), ..., (e_E, \hat{t}_E)\}$ be the set of truth priors, and let $\hat{t}_e$ be the prior for entity e. The detailed preprocessing steps are described in Algorithm 1. First, claimed values with relative errors or absolute errors above certain thresholds are treated as outliers. Second, a Gaussian distribution with the prior truth $\hat{t}_e$ as mean and variance of $\hat{\sigma}_e^2$ can measure the probability of observing each claimed value, which can be thresholded to detect outliers. This is equivalent to setting a threshold on the z-scores (how many standard deviations an observation is above or below the mean). The only issue is that the true $\hat{\sigma}_e$ is unknown at the beginning and outliers can make the estimate arbitrarily large. One possible solution we describe in Algorithm 1 is that we use relative errors and absolute errors to detect outliers first, then calculate $\hat{\sigma}_e$ without considering recognized outliers, and update $\hat{\sigma}_e$ every time new outliers are detected until there are no more outliers.

We are aware that more advanced outlier detection techniques exist, and they can be potentially deployed here, but thresholding Gaussian p-values is one of the most commonly used methods, and its effectiveness has been justified in the past. Since the focus of this work is not a novel outlier detection method, we just apply this simple approach with a small modification in the sense that we use

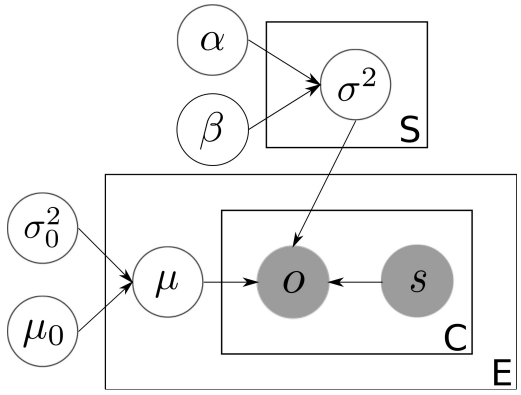

Figure 1: The probabilistic graphical model of GTM.

truth priors as the mean, and iteratively update the standard deviation. Our experiments show that it works quite well on our data. We also tested various strategies to process outliers after they are detected, and excluding them in truth-finding works best in practice compared with limiting their errors to a certain maximum value (Winsorising).

After the outliers are detected, we normalize all claimed values to its z-scores for the following truth-finding step. Specifically, for each claim $c \in \mathcal{C}$, $o_c = (v_c - \hat{t}_{e_c})/\hat{\sigma}_{e_c}$ is the transformed value. This is the most commonly used normalization step. In the previous section we have explained that the data scale and consensus level for each entity needs to be considered for adjusting the prior weight and assessing source quality, and such normalization would help reduce the various biases we discussed.

3.3 Model Details

Now we will explain the details of GTM (Figure 1 is the graphical representation of our model). Since GTM is a generative model, we will describe the conceptual generation process of each random variable, *i.e.*, the quality of each source, the truth of each entity and the observed value of each claim, and how they are dependent in our model.

3.3.1 Quality of Sources

For each source $s \in \mathcal{S}$, generate its quality σ_s^2 from a prior inverse Gamma distribution with hyper-parameter (α, β), where α is the shape parameter and β is the scale parameter:

$$\begin{aligned}\sigma_s^2 &\sim Inv\text{–}Gamma(\alpha, \beta) \\ &\sim (\sigma_s^2)^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma_s^2}\right).\end{aligned}$$

Here σ_s^2 is the variance parameter, and therefore lower values correspond to higher source quality, and higher values correspond to lower source quality respectively. The inverse Gamma distribution is utilized because it is the conjugate prior of the Gaussian distribution with σ_s^2 as variance, meaning the posterior of σ_s^2 is also an inverse Gamma distribution and the MAP inference is more efficient as a result. Its parameter α and β controls the prior belief about how source quality is distributed, *e.g.*, the expectation of σ_s^2 is given by $\frac{\beta}{\alpha}$. In practice, if prior knowledge about the quality of specific sources is available, corresponding hyper-parameters can be adjusted accordingly; otherwise, we could simply use the same prior for all sources, which has the effect of smoothing that alleviates bias caused by small volume data, *e.g.*, some sources may make very few claims and therefore the MLE estimation of its quality may not be accurate.

3.3.2 Truth of Entities

For each entity $e \in \mathcal{E}$, generate its truth μ_e from a prior Gaussian distribution with μ_0 mean and σ_0^2 variance:

$$\begin{aligned}\mu_e &\sim Gaussian(\mu_0, \sigma_0^2) \\ &\sim \exp\left(-\frac{(\mu_e - \mu_0)^2}{2\sigma_0^2}\right).\end{aligned}$$

Here μ_0 controls the prior belief about the location of the truth and σ_0^2 controls the weight of the prior. Since we have already normalized all claims to there z-scores based on $\hat{\mathcal{T}}$, we should also use the standard Gaussian as the prior here by setting $\mu_0 = 0$ and $\sigma_0^2 = 1$, although σ_0^2 can still be adjusted to reflect how much we believe $\hat{\mathcal{T}}$ is correct.

3.3.3 Observation of Claims

For each claim c of entity e, *i.e.*, $c \in \mathcal{C}_e$, denote its source as s_c, which is an observed dummy index variable that we use to select the corresponding source quality. We generate o_c, the normalized claimed value of c, from a Gaussian distribution with the truth of e as mean, and the variance parameter of s_c as its variance:

$$\begin{aligned}o_c &\sim Gaussian(\mu_e, \sigma_{s_c}^2) \\ &\sim \sigma_{s_c}^{-1} \exp\left(-\frac{(o_c - \mu_e)^2}{2\sigma_{s_c}^2}\right).\end{aligned}$$

Here truth of entities and quality of sources collectively controls the probability of observing each claim, which aligns with our intuitions that claims farther from the truth are less likely to be observed, and claims made by low quality sources are more likely to deviate from the truth.

3.4 Inference

Given the construction of GTM, the complete likelihood of observed data and unknown parameters given the hyper-parameters can be written as:

$$p(\boldsymbol{o}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 | \mu_0, \sigma_0^2, \alpha, \beta) = \prod_{s \in \mathcal{S}} p(\sigma_s^2 | \alpha, \beta) \times \prod_{e \in \mathcal{E}} \left(p(\mu_e | \mu_0, \sigma_0^2) \prod_{c \in \mathcal{C}_e} p(o_c | \mu_e, \sigma_e^2) \right).$$

Then truth finding is equivalent to searching for optimal truth estimates that maximize the joint probability, *i.e.*, get the *maximum a posterior* (MAP) estimate for $\boldsymbol{\mu}$:

$$\hat{\boldsymbol{\mu}}_{MAP} = \arg\max_{\boldsymbol{\mu}} \int p(\boldsymbol{o}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 | \mu_0, \sigma_0^2, \alpha, \beta) \mathrm{d}\boldsymbol{\sigma}^2 .$$

Let $\hat{\mu}_e$ be the MAP estimate for entity e output by GTM, then we transform it back to its actual value before normalization: $\hat{t} + \hat{\mu}_e \hat{\sigma}_e$, and predict it as the truth. If we know the truth is always claimed by at least one source, we can predict claimed values that are closest to it as the truth, and if there is a tie, the average of tied values can be taken as output.

Essentially various algorithms can be deployed for the MAP inference, such as EM [8], Gibbs Sampling [8], *etc.* Next, we will briefly describe an EM algorithm that iteratively computes truth $\boldsymbol{\mu}$ and source quality $\boldsymbol{\sigma}^2$, which further uncovers how they are related in our model.

Optimizing the likelihood is equivalent to optimizing its logarithm form, which is given by:

$$\begin{aligned} L =& \log p(\boldsymbol{o}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 | \mu_0, \sigma_0^2, \alpha, \beta) \\ \sim& -\sum_{s \in \mathcal{S}} \left(2(\alpha+1)\log\sigma_s + \frac{\beta}{\sigma_s^2} \right) - \sum_{e \in \mathcal{E}} \frac{(\mu_e - \mu_0)^2}{2\sigma_0^2} - \\ & - \sum_{e \in \mathcal{E}} \sum_{c \in \mathcal{C}_e} \left(\log\sigma_{s_c} + \frac{(o_c - \mu_e)^2}{2\sigma_{s_c}^2} \right). \end{aligned}$$

Then, in the E step, we assume $\boldsymbol{\sigma}^2$ is given, and achieve the optimal truth μ_e by solving $\frac{\partial L}{\partial \mu_e}$, which is:

$$\hat{\mu}_e = \frac{\frac{\mu_0}{\sigma_0^2} + \sum_{c \in \mathcal{C}_e} \frac{o_c}{\sigma_{s_c}^2}}{\frac{1}{\sigma_0^2} + \sum_{c \in \mathcal{C}_e} \frac{1}{\sigma_{s_c}^2}} = \frac{\frac{\mu_0}{\sigma_0^2} + \sum_{c \in \mathcal{C}_e} \frac{v_c - \hat{t}_e}{\sigma_{s_c}^2 \hat{\sigma}_e}}{\frac{1}{\sigma_0^2} + \sum_{c \in \mathcal{C}_e} \frac{1}{\sigma_{s_c}^2}}. \quad (1)$$

We can see in the above equation that more accurate sources have higher weights in deciding the truth.

Conversely, if truth $\boldsymbol{\mu}$ is given, we can get the optimal estimate for the quality of each source s by solving $\frac{\partial L}{\partial \sigma_s^2}$. Let $\mathcal{C}_s$ be all the claims made by s, we have the M step:

$$\hat{\sigma}_s^2 = \frac{2\beta + \sum_{c \in \mathcal{C}_s} (o_c - \mu_{e_c})^2}{2(\alpha+1) + |C_s|} = \frac{2\beta + \sum_{c \in \mathcal{C}_s} \left(\frac{v_c - \hat{t}_{e_c}}{\hat{\sigma}_{e_c}} - \mu_{e_c} \right)^2}{2(\alpha+1) + |C_s|}. \quad (2)$$

We can see source quality is estimated by how close the source's claims are to the truth, discounted by the variance of all claimed values for each entity and smoothed by prior parameters.

During such iterative computation of truth and source quality, the log likelihood will always increase in each iteration, and therefore a local maximum is guaranteed to be reached. The truth of each entity can be initialized to 0 at the very beginning, indicating our belief that the actual truth should be close to the prior.

4. EXPERIMENTS

In this section, we report the experimental results of how GTM performs on two real world numerical datasets compared with state-of-the-art algorithms.

4.1 Setup

4.1.1 Datasets

We use the following two datasets originally published in [10] in our experiments:

Population Dataset. This data is a sample of Wikipedia edit history of city population in given years, consisting 4119 claims about 1148 population entities from 2415 sources. 274 city-year pairs are randomly sampled and labeled with the true population.

Bio Dataset. This data is another Wikipedia edit history sample about people's dates of birth or death. 607819 sources make 1372066 claims about 9924 date entity. 1912 dates are verified and labeled as truth.

Several preprocessing steps have been applied on both datasets. First, since the data is Wikipedia edit history, there could be multiple versions of claims made by the same source about the same entity ordered by their timestamps. Considering sources may realize their mistakes and update their claims, we only keep the most recent claim of each source so that historical errors are not punished and estimated source quality is more accurate. We have verified that this step can improve the effectiveness of all the methods we compare on both datasets.

Second, we remove trivial records where all claims are the same. We thought this would make the problem more challenging, but our experiments show that it actually improved the effectiveness of all truth-finding methods on conflicting records in both datasets. The reason could be that algorithms may give sources contributing to the non-conflicting records more credit than what they actually deserve, since truth of these records may be easier to get. Another non-conflicting case is that there is only one source making a claim, which also should not contribute to the source's quality since it is uncertain if the claim is indeed true due to lack of evidence. Additionally, the evaluation should also only involve conflicting records so that the measurement will be more accurate on reflecting the actual performance of truth finding methods.

4.1.2 Compared Algorithms

We compare the effectiveness and efficiency of our Gaussian Truth Model (GTM) with several previous methods together with treating the most frequent value, the average, or the median as truth. We briefly summarize them as follows, and refer the reader to the original publications for details. In GTM, the claimed value closest to the estimated truth is output as the final prediction. If tie happens in all methods, the average of tied values is taken as the final output.

Voting. For each entity, output the value that is claimed by the most sources.

Median. The median of all claims is predicted as truth.

Average. The average of all claims is predicted as truth.

LTM [15]. A Bayesian probabilistic approach that focuses on categorical input and models the truth as latent Bernoulli variables. It assumes being true or not for different claimed values of the same entity are independent and multiple values can be true, which does not fit our problem setting. However, we still include this approach for comparison.

TruthFinder [13]. For each claimed value calculate the probability that at least one supporting claim is correct using the precision of sources.

3-Estimates [6]. A method that considers the difficulty of getting the truth when calculating source quality

Investment [10, 11]. Each source uniformly distributes its credits to its claims, and gains credits back from the confidence of those claims. Voting information is used as prior in this method as indicated in [10].

Parameters for the algorithms we compare are set according to the optimal settings suggested by their authors. For our method, on the population data the source quality prior is set as ($\alpha = 10, \beta = 10$); truth prior is set as ($\mu_0 = 0$, $\sigma_0^2 = 1$), and the results of TruthFinder are leveraged as the initial truth guess, since many entities have very few claims in this dataset and therefore Voting is not very accurate. On the bio data, the source quality prior is ($\alpha = 1, \beta = 100$), incorporating our belief that the quality of sources with small volume claims is lower on this data, which we learn from our observation. The truth prior is same: ($\mu_0 = 0, \sigma_0^2 = 1$); and Voting is used to give the initial estimates of the truth, since there are more claims for each entity, Voting tends to achieve better performance.

The default thresholds for the outlier detection step in our experiments are: on population data, threshold δ_0 on relative errors is 0.9, threshold δ_1 on absolute errors is not set, threshold δ_2 on z-scores

Table 2: Inference results per dataset and per method.

	Results on the population data		*Results on the bio data*	
	MAE	*RMSE*	*MAE*	*RMSE*
GTM	**1498.59**	**8339.99**	**228.19**	**4831.53**
3-Estimates	1640.83 (+9.49%)	8822.50 (+5.78%)	237.35 (+4.01%)	4847.80 (+0.33%)
TruthFinder	1633.60 (+9.00%)	8824.09 (+5.80%)	660.66 (+189.51%)	6959.72 (+44.04%)
Investment	1787.65 (+19.28%)	9358.80 (+12.21%)	3858.90 (+1591.04%)	26237.65 (+443.05%)
LTM	3040.90 (+102.91%)	12865.52 (+54.26%)	396.78 (+73.87%)	5837.66 (+20.82%)
Voting	10327.20 (+589.12%)	126217.98 (+1413.40%)	237.35 (+4.01%)	4847.80 (+0.33%)
Median	10241.81 (+583.42%)	126198.86 (+1413.17%)	244.04 (+6.94%)	4854.90(+0.48%)
Average	10368.54 (+591.88%)	126199.76 (+1413.18%)	253.41 (+11.05%)	4860.28 (+0.59%)

is 2.5; on bio data, only threshold δ_2 is set to be 10. Outlier detection achieves 98.46% precision (2 wrong outliers), 35.95% recall on population and 99.73% precision (4 wrong outliers), 49.10% recall on dates, where precision is defined as how much percentage of unique entity value pairs that are detected are indeed false, and recall is defined as how much percentage of unique false entity value pairs are output by outlier detection. We have investigated the wrong outliers and found those 6 entities either have very few claims or the truth is indeed quite far from the range claimed by most sources, so it is very difficult to get them right.

4.2 Experimental Results

4.2.1 Effectiveness of Truth Finding

To evaluate the effectiveness of truth-finding on numerical data, the evaluation measures should reflect how close predictions are to the ground truth in general rather than how often predictions are exactly correct. Therefore, we use Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) for evaluation, with the latter giving more penalty to larger errors.

Detailed results on both the population data and the bio data are reported in Table 2, which demonstrates that GTM outperforms state-of-the-art approaches on both datasets. All methods are executed on the same data with outliers detected and removed.

On the population data, there are many entities with very few claims, which makes Voting, Median and Average not stable. In comparison, 3-Estimates, TruthFinder and Investment can achieve better performance by considering source quality. And GTM can make further improvement in terms of both MAE and RMSE by leveraging the characteristics of numerical data.

On the bio data, there are much more claims for each entity in average, which makes methods that do not even consider source quality, such as Voting, Median and Average, much stable and accurate comparing with the population dataset. However, it also causes decreased performance of TruthFinder and Investment. The reason is that TruthFinder models the confidence of each value as the probability that at least one claim supporting the value is correct, based on the precision of each source. With numerous supporting claims, many claimed values can get a probability close to 1, which makes the method fail. Similar reasons also make Investment fail. Among all the methods, GTM has the best performance in terms of MAE and RMSE, although RMSE is not significantly higher than Voting, Median and Average, which may be due to RMSE is dominated by some difficult records that all methods make mistakes on.

One both datasets, the performance of LTM is significantly worse than the best methods. This is expected because the basic assumption of LTM that multiple claims can be true contradicts with the numerical setting. Therefore, in practice GTM and LTM should be properly deployed based on the characteristics of data.

5. RELATED WORK

The truth-finding problem has been extensively studied. [13] was the first to propose an algorithm that iteratively computes truth and source quality. [10] followed up and developed a few heuristic algorithms and integrated truth-finding with integer programming to enforce certain constraints on truth data. [11] further generalized algorithms in [10] so that certain background information such as uncertainty of input records can be incorporated. [6] proposed that the difficulty of getting the truth should be considered in computing source quality, with the consideration that sources should not gain too much credit from easy records. Recently, [15] proposed a Bayesian probabilistic approach that can naturally support multiple true values for each entity. Although these methods could be applied on numerical data, none of them fully leverage the characteristics of numerical data, which is why our proposed method can outperform these state-of-the-art methods in our experiments.

Past work also focuses on other aspects in truth-finding or general data integration. The source copying issue was examined in [2–5, 12]. With the copying relationship between sources detected, the true support for each record can be better estimated. Dynamic updates of the data are considered in [5, 9]. [1] discusses predicting price history from multiple web offers, but it focuses on the time series aspect instead of quality of sources. [14] models truth finding as a semi-supervised problem and utilizes regularization to enforce that similar claims should have similar confidence of being true. [7] focuses on integrating knowledge bases.

6. CONCLUSIONS

In this paper, we are the first to propose a Bayesian probabilistic model called the Gaussian Truth Model to solve the critically important problem of truth finding on numerical data. We identify the major challenge of this problem is how the truth-finding method can adapt to the characteristics of numerical data for improving the effectiveness. Based on our intuitions on how truths, source quality and claimed values should be dependent on each other, we leverage Bayesian graphical models to design the generation process of all the variables such that our intuitions about their dependencies can be modeled in a principled way. Experiments on two real world datasets demonstrate that our new method is more effective than state-of-the-art methods.

7. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable comments and suggestions. The work was supported in part by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA).

8. REFERENCES

[1] R. Agrawal and S. Ieong. Aggregating web offers to determine product prices. In *KDD*, 2012.

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

[3] X. L. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava. Global detection of complex copying relationships between sources. *PVLDB*, 3(1):1358–1369, 2010.

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

[5] X. L. Dong, L. Berti-Equille, and D. Srivastava. Truth discovery and copying detection in a dynamic world. *PVLDB*, 2(1):562–573, 2009.

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

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

[8] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[9] A. Pal, V. Rastogi, A. Machanavajjhala, and P. Bohannon. Information integration over time in unreliable and uncertain environments. In *WWW*, pages 789–798, 2012.

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

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

[12] A. D. Sarma, X. L. Dong, and A. Y. Halevy. Data integration with dependent sources. In *EDBT*, pages 401–412, 2011.

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

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

[15] 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.