Online Truth Discovery on Time Series Data

Liuyi Yao * Lu Su* Qi Li † Yaliang Li‡ Fenglong Ma* Jing Gao* Aidong Zhang*

Abstract

Truth discovery, with the goal of inferring true information from massive data through aggregating the information from multiple data sources, has attracted significant attention in recent years. It has demonstrated great advantages in real applications since it can automatically learn the reliability degrees of the data sources without supervision and in turn helps to find more reliable information. In many applications, however, the data may arrive in a stream and present various temporal patterns. Unfortunately, there is no existing truth discovery work that can handle such time series data. To tackle this challenge, we propose a novel online truth discovery framework that incorporates the predictions on the time series data into the truth estimation process. By jointly considering the multi-source information and the temporal patterns of the time series data, the proposed framework can improve the accuracy of the truth discovery results as well as the time series prediction. The effectiveness of the proposed framework is validated on both synthetic and real-world datasets.

Keywords: Truth discovery, Streaming data, Time series

1 Introduction

In the big data era, effectively managing databases is extremely important due to the redundant data generated and stored continuously. Multiple sources provide data for the same object, where the sources can be websites, sensors, and human workers. Conflicts among them are inevitable due to various reasons, such as the quality of the sensors and the knowledge of human workers. Facing the daunting scale of the data, it is hard for people to judge which piece of information is accurate or which data source is reliable. Therefore, unsupervised approaches that can automatically find trustworthy information (which is usually referred to as *truth*) from the noisy and conflicting data are much desired. Among them, aggregation can be a good approach to resolve the conflicts and correct errors, since it can cancel out the errors made by individual data sources.

The most straightforward aggregation approach is to conduct voting or averaging on the multi-sourced data. This approach is simple and efficient. However, it has an obvious drawback: The sources are assumed to be equally reliable, which is usually untrue in real life. From our daily experience, we know that some sources are more reliable than the others. If such information can be captured, the aggregated results may be significantly improved. However, prior knowledge of reliability is not available, so it has to be inferred from data. Based on this idea, *truth discovery* methods stand out from the aggregation approaches thanks to the incorporation of source reliability estimation. In truth discovery methods, the estimation of source reliability and the inference of truth are tightly combined so that both the source reliability and the truth can be learned from the data in an unsupervised manner. If a source often provides trustworthy information, it will be assigned with high reliability; and in turn, if one piece of information is claimed by many reliable sources, it will be regarded as the final truth.

Traditional truth discovery methods [1, 2, 3, 4, 5] conduct iterative procedures of source reliability estimation and truth inference usually on static data. In recent years, online truth discovery methods [6, 7, 8] are proposed to handle streaming data. However, those methods either ignore the temporal patterns of evolving truths or simply assume that the truth values at consecutive time slots are similar. This assumption is only valid for a small portion of real-world applications. In many applications, *time series data*, such as the temperature, precipitation, and traffic volume data, are generated in a streaming manner. Temporal recurrences of similar phenomenon patterns, or *seasonal trends* [9], are commonly observed from such data, potentially at various different time scales (e.g., temperature is usually higher at daytime and lower at night, and also higher in the summer and lower in winter; traffic rush hours are observed each day, and also follow the weekdays-vs-weekend patterns). These patterns are helpful information for the inference of truth. Therefore, *how to model and incorporate the seasonal trends of the evolving truths in truth discovery* is the question that will be answered in this paper.

In this paper, we present **OTD**, an **O**nline **T**ruth **D**iscovery framework for time series data. OTD is a novel optimization framework that combines two components, namely the *truth discovery component* and the *time series analysis component*. As a result, these two components enhance each other. The truth discovery is guided by the predictions from the time series analysis, and the time series model is refined by the truth discovery results. In the truth

*SUNY Buffalo, {liuyiyao, lusu, fenglong, jing, azhang}@buffalo.edu
†University of Illinois at Urbana-Champaign, qili5@illinois.edu
‡Baidu Research Big Data Lab, yaliangli@baidu.com

discovery component, a summation of weighted errors is formulated as the objective value so that the sources whose claims are closer to the estimated truths will have higher weights. In the time series analysis component, Seasonal ARIMA (SARIMA) model [9] is used to capture the diverse seasonal trends of the time series data. The two components are linked together by the estimated truths: the truths should be close to the claims from reliable sources, as well as the predictions made by the time series model. The balance of the two components is carefully controlled so that the global error can be minimized. The proposed solution to this optimization problem is an online algorithm that does not need to store all the historical data. To test the effectiveness of the proposed methods, we conduct various types of experiments on both real-world and simulated datasets. The experimental results clearly demonstrate the improvement of the accuracy on the truth estimations as well as the time series predictions.

To summarize, we make the following contributions:

(1) To the best of our knowledge, the proposed OTD framework is the first truth discovery approach that can be applied the time series data, which considers both the smoothly evolving truths and the ones with seasonal trends.

(2) The truth discovery component in the proposed OTD enhances the time series modeling, and in turn, the time series modeling can help the truth discovery component improve the accuracy of truth estimation.

(3) The proposed OTD is built upon an online algorithm so that it preserves both efficiency and accuracy.

(4) We validate OTD on both synthetic and real-world datasets. The experiment results clearly demonstrate the effectiveness of the proposed method in finding reliable sources and inferring trustworthy information.

2 Methodology

The proposed OTD framework is formally presented in this section. We first describe the truth discovery problem settings for the streaming data, and then formulate the problem as an optimization framework. Finally, an online solution is provided.

2.1 Problem Settings Input. Suppose there are totally O objects that we are interested in. At each timestamp $t \in \{1, 2, \cdots, T\}$, there are some sources who provide claims on those objects. We denote the claim from the source s at timestamp t on object i as $x^s_{i,t}$, and the set of all claims at t as $\mathcal{X}_t = \left\{x^s_{i,t}\right\}^{O,S}_{i=1,s=1}$, where S is the number of total sources.

Output. Our goal is to find the most trustworthy information for each object at each timestamp, i.e., $\left\{x^*_{i,t}\right\}^{O,T}_{i=1,t=1}$, where $x^*_{i,t}$ is defined as the truth of object i at timestamp t.

In addition, $\mathcal{X}^*_t = \left\{x^*_{i,t}\right\}^{O}_{i=1}$ denotes the set of aggregation results at timestamp t, and OTD also estimates the source reliability degrees, i.e., the source weights. We denote the weight of the source s as w_s, and the set of all source weights as $\mathcal{W}$. A high source weight indicates that the source is reliable.

2.2 OTD Framework. The key idea of the proposed OTD framework is to incorporate the temporal patterns of the streaming data into the truth discovery process. The estimated truths should be close to the claims from reliable sources, and at the same time, follow the evolving pattern learned from the history. By doing so we can find more accurate truths.

OTD contains two components. At each timestamp, we first mine the patterns of the truth evolution by learning a time series model from the historic data and predict object truths for the current timestamp. Then we combine the predictions with the claims from sources to estimate object truths and source weights.

Mathematically, we formulate the truth discovery on streaming data with various patterns as to minimize the following overall loss:

(2.1)
$$\min_{\mathcal{W},\{\mathcal{X}^*_t\}^T_{t=1}} = \sum_{t=1}^{T}\left(\frac{1}{2}\sum_{s=1}^{S} w_s \sum_{i\in\mathcal{C}^s_t}(x^s_{i,t} - x^*_{i,t})^2 - \sum_{s=1}^{S} c^s_t \log(w_s) + \lambda\sum_{i=1}^{O}\frac{1}{R_{i,t}} L_\delta(x^*_{i,t}, \hat{x}^*_{i,t})\right),$$

where $\mathcal{W}$ is the set of all source weights and w_s is the weight of source s; $\mathcal{C}^s_t$ is the index set of objects on which source s made claims at timestamp t; $x^s_{i,t}$ is the claim of source s on object i at timestamp t; $x^*_{i,t}$ is the **estimated** truth of object i at timestamp t; c^s_t is the number of claims provided by s at t; $\hat{x}^*_{i,t}$ is the **predicted** value of object i's truth at timestamp t, and $R_{i,t}$ is the error degree of $\hat{x}^*_{i,t}$. Details about $\hat{x}^*_{i,t}$ and $R_{i,t}$ are introduced in Section 2.2.1. L_δ denotes the Huber Loss:

$$L_\delta(a,b) = \begin{cases} \frac{1}{2}(a-b)^2 & \text{if } |a-b| < \delta, \\ \delta|a-b| - \frac{1}{2}\delta^2 & \text{otherwise,} \end{cases}$$

where δ is a constant.

The first term, $\sum_{s=1}^{S} w_s \sum_{i\in\mathcal{C}^s_t}(x^s_{i,t} - x^*_{i,t})^2$, in the loss function measures the weighted L_2 distance from source claims to the estimated truths. The second term $\sum_{s=1}^{S} c^s_t \log(w_s)$ is a constraint to ensure that the source weights are positive. These two terms together will give a big penalty if the estimated truths are far from the claims given by the sources of high weights. As a result, the estimated truths will be close to the claims from reliable sources, and a source will get a high weight if its claims are close to the estimated truths.

The third term is the distance between the estimated truth $x^*_{i,t}$ and its predicted truth $\hat{x}^*_{i,t}$. The predicted truth $\hat{x}^*_{i,t}$ can be obtained from mining the temporal patterns of the

truth from the historic streaming data. $R_{i,t}$ denotes the error degree of the predicted truth. The smaller $R_{i,t}$, the more accurate the predicted truth. The details of $R_{i,t}$ are illustrated in 2.2.1. Intuitively, *by minimizing the difference between the estimated and predicted truths, we can incorporate the evolving patterns of the objects into the truth estimation*. To achieve this, we use the Huber loss to measure the distance between estimated truth and predicted truth. Huber loss is a hybrid of squared error (for relatively small errors) and absolute error (for relatively large ones). This loss function is used because it is differentiable and can give robust estimations [10]. With Huber loss function, we can prevent the estimated truths from being affected by extremely inaccurate predicted truths.

Next, we first introduce a time series model to predict object truths in Section 2.2.1, and then give detailed solution to the above overall optimization problem in Section 2.2.2.

2.2.1 Mining the Temporal Patterns of the Truths. At timestamp t, given the previous truth information $\{x^*_{i,j}\}_{j=1}^{t-1}$ of object i, our goal is to 1) first mine the truth evolving patterns, and 2) give a prediction on the current truth $x^*_{i,t}$. SARIMA [9] is used to model the evolving pattern, as it can capture seasonal trend, which denotes the phenomenon that similar patterns appear repeatedly over some periods. The seasonal trend is common in many real world scenarios. For example, the temperature of weather presents seasonal trend over both a day (high at daytime and low at night) and a year (hot in summer and cold in winter). Another example is the volume of traffic, which presents seasonal trend over both a day (rush hour in the morning and evening) and a week (weekdays and weekend).

To fit the streaming data, an online algorithm is needed to estimate coefficients in SARIMA. In order to simplify this problem, motivated by the online ARIMA algorithms in [11, 12, 13], we approximate SARIMA $(p,d,q)\times(P,D,Q)_E$ by ARIMA$(M+p+EP, d+DE, 0)$ with fixed $M \in \mathbb{N}$:

$$(1-\mathcal{B})^d(1-\mathcal{B}_E)^D x^*_{i,t} = \sum_{k=1}^{M+p+EP} \gamma_k (1-\mathcal{B})^d(1-\mathcal{B}_E)^D x^*_{i,t-k} + z_t,$$

where:

• $\mathcal{B}$ and $\mathcal{B}_E$ denote the backward shift operators:

$$\mathcal{B}x^*_{i,t} = x^*_{i,t-1};\ \mathcal{B}^k x^*_{i,t} = x^*_{i,t-k};\ \mathcal{B}_E x^*_{i,t} = x^*_{i,t-E};$$
$$\mathcal{B}^k_E x^*_{i,t} = x^*_{i,t-kE};\ \mathcal{B}^l\mathcal{B}^k_E x^*_{i,t} = x^*_{i,t-kE-l}; \mathcal{B}^l\mathcal{B}^k_E = \mathcal{B}^k_E\mathcal{B}^l.$$

• d and D are the regular difference order and seasonal difference order, respectively. The result of applying regular difference on $\{x^*_{i,j}\}_{j=1}^{t-1}$ d times is $\{(1-\mathcal{B})^d x^*_{i,j}\}_{j=d+1}^{t-1}$; and the result of applying seasonal regular difference on $\{x^*_{i,j}\}_{j=1}^{t-1}$ D times is $\{(1-\mathcal{B}_E)^D x^*_{i,j}\}_{j=d+1}^{t-1}$.

• γ_k is the k-th entry of approximated ARIMA model parameter γ ($\gamma \in \mathbb{R}^{M+p+EP}$).

• E is the period. For daily observations, like the traffic volume, which has weekly trend, E usually is 7; For monthly observations, like monthly average temperature, E usually is 12 (12 months in 1 year).

• p, P, q, Q are the orders of Auto-Regressive (AR) process, Seasonal AR process, Moving-Average (MA) process, Seasonal MA process in SARIMA separately.

• z_t denotes white noise.

By approximation, we only need to estimate an $(M+p+EP)$-dimensional coefficient vector γ. Online gradient decent is adopted to estimate coefficient vector γ. To better describe the coefficient estimation, the following notations are introduced. $\mathcal{K}$ is the set of candidate coefficient vectors: $\mathcal{K} = \{\gamma \in \mathbb{R}^{M+p+EP}, |\gamma_k| \le g, k = 1, ..., M\}$, where g is a positive constant and γ_k is the k-th element in the coefficient vector γ; Π denotes the projection operator: $\Pi_{\mathcal{K}}(c) = \arg\min_{y\in\mathcal{K}} \|c-y\|_2$, where c is a constant vector.

We first initialize γ as γ^0, where $\gamma^0 \in \mathcal{K}$. At timestamp t, we get the truth at previous timestamp $x^*_{i,t-1}$. Thus we can update the coefficients by the gradient descent based on the prediction error on $x^*_{i,t-1}$. The prediction error at timestamp t with respect to γ^{t-1} (the value of γ at timestamp $t-1$) is:

$$l^M_{i,t}(\gamma^{t-1}) = \Bigg[\sum_{k=1}^{M+p+EP} \gamma_k^{t-1}(1-\mathcal{B})^d(1-\mathcal{B}_E)^D x^*_{i,t-1-k} + \sum_{k=1}^{d+DE}\binom{d+DE}{k}(-1)^k x^*_{i,t-k-1} - x^*_{i,t-1}\Bigg]^2. \tag{2.2}$$

We use the derivative of $l^M_{i,t}(\gamma^{t-1})$ on γ^{t-1}, to update γ as:

$$\gamma^t = \Pi_{\mathcal{K}}\left(\gamma^{t-1} - \frac{1}{\eta}\nabla l^M_{i,t}(\gamma^{t-1})\right), \tag{2.3}$$

where η is the learning rate.

Therefore, the $\hat{x}^*_{i,t}$ in the third term of the loss function Eqn. (2.1) can be calculated with the updated parameters as:

$$\hat{x}^*_{i,t} = \sum_{k=1}^{M+p+EP} \gamma_k^t(1-\mathcal{B})^d(1-\mathcal{B}_E)^D x^*_{i,t-k} + \sum_{k=1}^{d+DE}\binom{d+DE}{k}(-1)^k x^*_{i,t-k}. \tag{2.4}$$

Algorithm 1 summaries the detailed steps of mining truth evolving patterns and predicting the truths at the current timestamp, given predefined parameters p, P, E, d, D, M and η. As we can see, from the first two lines, the algorithm conducts parameter estimation, and in line 3, the algorithm predicts the truth at the current timestamp.

Prediction Regularization. $\lambda\sum_{i=1}^{O}\frac{1}{R_{i,t}}L_\delta(x^*_{i,t}, \hat{x}^*_{i,t})$, which is the third term in the loss function Eqn. (2.1), is the prediction regularization term. In this term, λ can be viewed as the global control parameter on the overall importance

Algorithm 1: Online Truth Prediction

Input : Coefficient from last timestamp γ^{t-1}; Historical truth $\{\mathcal{X}_j^*\}_{j=t-1-M-p-EP}^{t-1}$;

Output: Current timestamp prediction $\hat{x}_{i,t}^*$; Coefficient γ^t; Square loss $l_{i,t}^M(\gamma^{t-1})$;

1 Calculate square loss $l_{i,t}^M(\gamma^{t-1})$ and $\nabla l_{i,t}^M(\gamma^{t-1})$ using Eqn. (2.2);
2 Update γ from γ^{t-1} to γ^t using Eqn. (2.3);
3 Get the prediction $\hat{x}_{i,t}^*$ using Eqn. (2.4);
4 **Return** $\hat{x}_{i,t}^*$, $l_{i,t}^M(\gamma^{t-1})$ and γ^t.

level of the truth predictions in the truth estimation procedure. $R_{i,t}$ measures the error degree of predictions, i.e., the fitness of the SARIMA model. If the model fits the data well, we will have a high confidence of the pattern of the evolving truths captured by the time series model, so the truth prediction is more accurate and should be trusted. On the other hand, if the SARIMA model does not fit the data well, we will have little confidence on its predictions. $R_{i,t}$ is defined as the expected root mean square error of the truth prediction error: $R_{i,t} = \sqrt{\sum_{k=1}^{t-1} l_{i,k}^M(\gamma^{k-1})/(t-1)}$, where $l_{i,k}^M(\gamma^{k-1})$ is defined in Eqn. (2.2). More penalty will be given if the estimated truth is far from the predicted truth of high confidence. As a result, the estimated $x_{i,t}^*$ will be close to $\hat{x}_{i,t}^*$ when $\frac{\lambda}{R_{i,t}}$ is large.

2.2.2 Weight and Truth Computation. We propose to solve the optimization problem (i.e., Eqn. (2.1)) using block coordinate descent method [14]. The basic idea is as follows: at each timestamp, we update the values of object truths and source weights (i.e., w_s) alternatively and separately:

Truth Update: At timestamp t, we first fix source weight $\mathcal{W}$ and solve the optimization problem with respect to only the estimated truth. By setting the partial derivative to 0, the update of estimated truths is as follows.

• If the predicted truths are not available:

$$x_{i,t}^* = \frac{\sum_{s=1}^S w_s x_{i,t}^s}{\sum_{s=1}^S w_s}. \tag{2.5}$$

• If the predicted truths are available:

If $|\frac{\sum_{s=1}^S w_s x_{i,t}^s + \frac{\lambda}{R_{i,t}}\hat{x}_{i,t}^*}{\sum_{s=1}^S w_s + \frac{\lambda}{R_{i,t}}} - \hat{x}_{i,t}^*| < \delta$:

$$x_{i,t}^* = \frac{\sum_{s=1}^S w_s x_{i,t}^s + \frac{\lambda}{R_{i,t}}\hat{x}_{i,t}^*}{\sum_{s=1}^S w_s + \frac{\lambda}{R_{i,t}}}; \tag{2.6}$$

If $\frac{\sum_{s=1}^S w_s x_{i,t}^s - \frac{\lambda}{R_{i,t}}}{\sum_{s=1}^S w_s} - \hat{x}_{i,t}^* < -\delta$:

$$x_{i,t}^* = \frac{\sum_{s=1}^S w_s x_{i,t}^s - \frac{\lambda}{R_{i,t}}}{\sum_{s=1}^S w_s}; \tag{2.7}$$

If $\frac{\sum_{s=1}^S w_s x_{i,t}^s + \frac{\lambda}{R_{i,t}}}{\sum_{s=1}^S w_s} - \hat{x}_{i,t}^* > \delta$:

$$x_{i,t}^* = \frac{\sum_{s=1}^S w_s x_{i,t}^s + \frac{\lambda}{R_{i,t}}}{\sum_{s=1}^S w_s}. \tag{2.8}$$

From the above derivations, we can see that if the predicted truth is close to the estimated truth, then it should be included in the truth estimation (according to Eqn. (2.6)). On the other hand, if the predicted truth is far from the estimated truth, then it is not included in the truth estimation (according to Eqn. (2.7) and Eqn. (2.8)).

As can be seen, the above four equations comply with the basic principles of truth discovery, i.e., the source with a higher weight plays a more important role in the truth estimation.

Weight Update: Then we fix the estimated truths , and update source weights as follows:

$$w_s = \frac{\sum_{j=1}^t c_j^s}{\sum_{j=1}^t \sum_{i \in \mathcal{C}_j^s} (x_{i,j}^s - x_{i,j}^*)^2}. \tag{2.9}$$

From Eqn. (2.9), it can be seen that the source weight calculation follows the basic principle of truth discovery that if a source provides information far from the estimated truth, the weight should be low, and vice versa.

2.2.3 Summary. So far, we have described how to model truth evolving patterns, and how to make use of the predicted truths to estimate object truths as well as user weights. Here we summarize the overall flow of the proposed online truth discovery framework in the following steps:

Step I: Invoke Algorithm 1 to update the prediction model and predict the current truths.

Step II: Use the predicted truths and the weight information to update the estimated truths.

Step III: Use the estimated truth to update the source weights.

Note that the SARIMA model needs at least $(M + p + EP + d + D + 2)$ data to build, so when $t < M + p + EP + d + D + 2$, Step I is skipped.

The major contribution of the proposed framework lies in the combination of truth discovery with the mining of truth evolving patterns. This joint design can improve both the truth discovery results and the time series analysis. The proposed online algorithm is efficient and does not need to store massive historical data. Thus, it can be applied to a full spectrum of applications that involve the analysis of streaming time series data.

3 Experiments

In this section, we experimentally evaluate the proposed OTD framework on synthetic datasets and real-world dataset.

3.1 Experiment Setup. We first describe the experiment setups that ensure a fair comparison between the proposed method and various baseline methods.

3.1.1 Performance Measures. The following two performance metrics are adopted for the purpose of evaluation: (1) Mean of Absolute Error (MAE) measures the L_1-norm distance from the estimated truths to the ground truths; (2) Root Mean of Square Error (RMSE) measures the L_2-norm distance between the estimated truths to the ground truths. L_1-norm distance (MAE) penalizes more on small errors, while L_2-norm distance (RMSE) focuses more on big errors. They are complementary. For both MAE and RMSE, the lower the value, the better the performance.

3.1.2 Baseline Methods. We compare the proposed OTD framework with several baseline methods, including: Streaming truth discovery method: **DynaTD+All** [6]; Non-streaming truth discovery methods: **Truthfinder** [1], **Accusim** [2], **Investment** [15], **2-Estimates** and **3-Estimates** [16], **GTM** [17], **CRH** [18], and **CATD** [4].

Besides the above methods, we also include **Mean** and **Median** baseline methods, which take mean or median value of all the claimed values as estimated truths.

Online Truth Prediction: We only use historical truth information to estimate current truth without incorporating multi-sources' claims, i.e., at time stamp t, use Algorithm 1's output $\hat{x}^*_{i,t}$ as the estimated truth.

3.2 Experiments on Synthetic Data. This set of experiments on synthetic datasets are designed to demonstrate the benefits of the proposed OTD framework as follows. (1) OTD can deal with various types of time series data. (2) OTD can be applied to both fixed source reliability and dynamic source reliability scenarios. (3) In OTD, we combine the aggregation of multi-source claims and the prediction information from online time series model, which leads to the best performance.

3.2.1 Data Generation. To fulfill the above goals, we consider the following four different cases:

Case 1: *Unsmoothly evolving truth and fixed source reliability*. We generate the object truths with the following parameters: AR process coefficients $(0.7, -0.6, 0.4, -0.5, 0.3)$, MA process coefficients $(0.5, -0.3)$, order $d = 1$, and other parameters are set as 0. In order to embed seasonal trend into the unsmoothly evolving truth, we add $\sin$ function with period $pd = 5$, amplitude $amp = 4$ to the generated truths. Then four sources are simulated, with their error distributions $N(0, \sigma^2)$ set as $\sigma = 0.5, 1, 1.5, 2$ respectively.

Case 2: *Unsmoothly evolving truth and dynamic source reliability*. Object truths are generated with the following parameters: AR process coefficients $(0.6, -0.6, 0.4, -0.5, 0.3)$, MA process coefficients $(0.3, -0.2)$, order $d = 1$ and other parameters are set as 0. Then we add $\sin$ function with period $pd = 5$, amplitude $amp = 4$ to the generated truths. Five sources with dynamic reliability are simulated in two steps: for each timestamp, we first generate each source's error distribution parameter σ from $N(1, 0.5)$, then generate errors from $N(0, \sigma^2)$, and add the generated errors to the corresponding truths.

Case 3: *Smoothly evolving truth and fixed source reliability*. We generate object truths and simulate sources following the same procedure as in Case 1, except that AR process coefficients $(0.6, -0.3, 0.4, -0.6, 0.5)$, MA process coefficients $(0.3, -0.2)$, period $pd = 20$ and $amp = 0.5$. In this case, four sources are simulated, with their error distributions $N(0, \sigma^2)$ set as $\sigma = 0.8, 1, 1.5, 2$ respectively.

Case 4: *Smoothly evolving truth and dynamic source reliability*. Object truths and sources are generated using the same way as in Case 2, except the AR process coefficients $(0.6, -0.5, 0.4, -0.4, 0.3)$ and the period $pd = 10$.

3.2.2 Performance Comparison. The results of the proposed OTD and all the baseline methods are summarized in Table 1. From these results, we observe that OTD achieves the best performance in Case 1, 2, and 3 under both MAE and RMSE performance measures, and in Case 4, OTD achieves similar results to DynaTD+All. It confirms that OTD can handle various scenarios, including smoothly and unsmoothly evolving truths, dynamic and fixed source reliability. The detailed performance analysis are as follows:

Generally speaking, when source weights are fixed (Case 1 and Case 3), most of truth discovery methods give better performance than Mean and Median, due to the contribution of the source reliability estimation component in truth discovery. However, this advantage may become less obvious when source reliability changes, as it is difficult for non-streaming truth discovery methods to accurately estimate dynamic source reliability.

Within the non-streaming truth discovery methods, **GTM**, **CRH** and **CATD** give better performance than others as these three methods are designed for continuous (or heterogeneous) data type.

As mentioned above, **DynaTD+All** is a streaming truth discovery method that can handle smoothly evolving truths and dynamic source reliability. Thus, this method gives good performance for Case 3 and Case 4 (smoothly evolving truth cases), while its performance under Case 1 and Case 2 (unsmoothly evolving truth cases) is not satisfactory due to its strong smoothness assumption about the temporal patterns of evolving truths. As a comparison, the proposed OTD method relaxes the assumption about smoothness, and the online time series prediction can help OTD to capture both smooth and unsmooth patterns in object truths. Besides, in the scenarios of smoothly evolving truths (Case 3 and Case

Table 1: Performance Comparison on Synthetic Datasets

	Case 1		Case 2		Case 3		Case 4	
Method	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Mean	0.5755	0.7129	0.4339	0.5567	0.5975	0.7392	0.3912	0.4906
Median	0.5118	0.6540	0.3249	0.4816	0.5662	0.7159	0.4442	0.5649
TruthFinder	0.4717	0.6164	0.4286	0.6346	0.6299	0.7905	0.6021	0.7510
AccuSim	0.4700	0.6206	0.4256	0.6315	0.6388	0.7989	0.6321	0.7978
Investment	0.4031	0.4967	0.7570	1.1650	0.6430	0.7968	1.0890	1.3801
2-Estimates	1.5728	1.9663	0.7762	1.1812	1.5771	1.9703	1.1033	1.4067
3-Estimates	0.4031	0.4967	0.7570	1.1650	0.6430	0.7968	1.1422	1.4620
GTM	0.4655	0.5876	0.4677	0.6077	0.5346	0.6756	0.4256	0.5354
CRH	0.4070	0.5107	0.4573	0.6098	0.4887	0.6158	0.4562	0.5726
CATD	0.4099	0.5149	0.4565	0.6068	0.4914	0.6195	0.4711	0.5942
DynaTD+All	0.7120	0.8759	0.6048	0.7474	0.3698	0.4613	0.3213	0.3999
Online Prediction	0.4839	0.6266	0.4461	0.7689	0.4119	0.5314	0.4174	0.5290
OTD	**0.2921**	**0.3750**	**0.3128**	**0.4276**	**0.3341**	**0.4370**	**0.2869**	**0.3730**

4), OTD has better performance on fixed source reliability case (Case 3), while DynaTD+All performs slightly better on dynamic source reliability case (Case 4). The reason is that DynaTD+All and OTD have different ways to calculate source weights. DynaTD+All penalizes more on the errors that a source made at the timestamp close to the current timestamp, in other words, DynaTD+All calculates the *"local source weights"* [6]. Thus, for the dynamic source reliability case (Case 4), it can estimate source weights more accurately. In contrast, OTD uses all errors a source made to calculate the source weights, i.e., OTD calculates the *"global source weights"*. Therefore, OTD performs better in the fixed source reliability case (Case 3).

Although the online time series prediction method has the ability to capture various patterns in object truths, it fails to utilize the multi-source information. Thus, the performance of baseline method **Online Prediction** is also not satisfactory.

The proposed OTD framework combines the weighted aggregation results of multi-source claims and the prediction results of online time series prediction method. This strategy integrates the benefits from both multi-source data and time series prediction. This leads to great performance improvement. For example, in Case 3, the performance of OTD is 9.65% better than the best baseline DynaTD+All only using multi-sources claims under MAE measure, and 5.27% better under RMSE measure. Compared with online time series prediction, the performance of OTD is 18.89% better under MAE measure, and 17.76% better under RMSE measure.

3.3 Experiments on Real-World Data. In this section, we conduct experiments on two real-world datasets. This set of experiments demonstrates that: (1) the proposed OTD framework works superior in real-world scenarios, and (2) the combination of multi-source data and online prediction can lead to performance improvement even when the predictions are not good enough.

3.3.1 Data Collection. In this experiment, we use two real-world datasets: Weather Dataset and Pedestrian Count Dataset. The data collection procedure is as follows:

Weather Dataset. We collect weather forecast information (high temperature and low temperature) about 88 US cities from three platforms: Wunderground[1], HAM weather[2], and World Weather Online[3]. The data collection started on October 7, 2013, and ended on January 1, 2014, which leads to a dataset consisting of $1,873,978$ records. Meanwhile, true high and low temperature information is collected for the purpose of evaluation. In this dataset, the objects are the highest temperature and lowest temperature of 88 US cities. Sources are three weather forecast platforms.

Pedestrian Count Dataset. This data is published by Dublin City Council[4]. In this dataset, daily pedestrian counts of four streets (Capel Street, Henry Street, Mary Street and O'connel street clearys) in 2015 are recorded. In the real world, there are many sources that can provide pedestrian counts. For example, surveillance cameras, infrared beam counters, thermal imaging systems, the sensors on the traffic signals, and the number of smart phone connections to Wi-Fi hot spots. Different ways of pedestrian counting have different reliability levels. Since it is hard to collect the claims of the aforementioned six systems, instead, we use

[1] http://www.wunderground.com
[2] http://www.hamweather.com
[3] http://www.worldweatheronline.com
[4] https://data.gov.ie/dataset/pedestrian_footfall

Table 2: Performance Comparison on Real-World Dataset

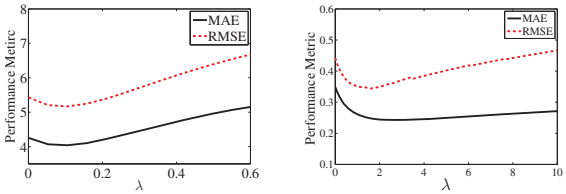
Method	Weather Dataset		Pedestrian Count	
	MAE	RMSE	MAE	RMSE
Mean	4.9240	6.3182	0.5134	0.6472
Median	4.6603	6.0908	0.5339	0.6747
TruthFinder	4.3899	5.8306	0.8164	1.0405
AccuSim	4.5684	6.0300	0.7198	0.9105
Investment	4.2980	5.7454	0.8164	1.0405
2-Estimates	4.2460	5.6353	1.6211	2.0274
3-Estimates	4.6233	6.2215	1.4645	1.8232
GTM	4.4230	5.6670	0.5214	0.6592
CRH	4.3021	5.6661	0.4934	0.6177
CATD	4.3921	5.6262	0.4661	0.5863
DynaTD+all	4.2442	5.4142	0.4213	0.5313
Online Prediction	7.3338	9.6724	0.3507	0.7130
OTD	**4.0378**	**5.1650**	**0.2797**	**0.4242**

Gaussian noise with different variances (σ^2) to simulate the error distributions of these sources. The variance of a source represents its reliability. The lower the variance, the higher the reliability. The variances are set as 1, 1.44, 1.96, 2.56, 3.24, 4, respectively. Thus, the claims from various sources are generated by adding different Gaussian noise to the ground truth. In this dataset, the objects are every day's pedestrian counts of four streets in 2015, and the sources are six simulated pseudo sources. In the following experiments, we set $\lambda = 0.1(2.6)$ and $\delta = 9(2)$ on the Weather (Pedestrian Count) Dataset.

3.3.2 Performance Comparison. Table 2 summarizes the results for OTD and all the baseline methods on the collected datasets. From these results, we have similar observations as shown on synthetic datasets: (1) Truth discovery methods have better performance compared with simple Mean and Median methods; (2) Streaming truth discovery method **DynaTD+all** achieves lower errors than non-streaming truth discovery methods; (3) The proposed OTD framework outperforms all the baseline methods under both MAE and RMSE measures. These observations confirm that OTD has the ability to capture complex patterns in real-world time series data.

Analysis on Online Prediction. From Table 2, we can observe that the accuracy of online prediction is not good on the Weather dataset. To explore the effect of online prediction on the proposed method, we vary the parameter λ, and the results are shown in Figure 1. For the Weather dataset, the best value for λ is around 0.1, a relatively small value.

As comparison, the best values for λ on the Pedestrian Count dataset is around 1.5 (refer to Figure 1b). This is because the accuracy of online time series prediction on the Weather dataset is not good enough, and we cannot rely too much on it.



(a) Weather Dataset: λ (b) Pedestrian Count: λ

Figure 1: The effect of parameter λ

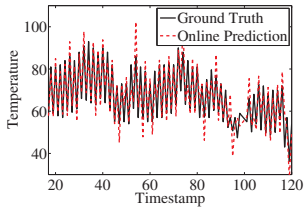


Figure 2: Prediction Results on Weather Dataset

However, from Table 2, we still observe the performance improvement achieved by OTD through incorporating online prediction information on the Weather dataset. To investigate the reason behind this, we plot the results of online prediction and ground truth information in Figure 2. Due to space limitation, we only show the plot for one randomly selected city. For other cities, similar observations can be made. Figure 2 shows that for some timestamps, the error of online prediction is big; while for some timestamps, the prediction is accurate. In OTD, there is a parameter $R_{i,t}$ to control the effect of online prediction on object i at timestamp t. $R_{i,t}$ can be treated as a local control parameter while λ is a global control parameter as it adjusts the effect of prediction results for all the objects at all the timestamps. Thus, when the prediction results are good at some timestamps for some objects, the local parameter $R_{i,t}$ will increase the effect of these prediction results, and vice versa. This is the reason that although the overall accuracy of online prediction on this real-world dataset is not good enough, it can still be helpful to improve the performance of the proposed method.

The most interesting finding is that *the performance of truth discovery and the online truth prediction can benefit each other*. In the above, we demonstrate that online truth prediction can help truth discovery. Now, we examine how truth discovery can help truth prediction. We randomly choose one city in the Weather dataset, and apply Online Truth Prediction algorithm on its previously estimated truths to get the prediction of the current truths. Then, we ran-

Table 3: Prediction Performance on Estimate Truths and Source Claims

Predicted Truths	MAE	RMSE
Predicted Truths from Historical Estimated Truths	6.1881	8.4418
Predicted Truths from a Single Source's Claims	6.6682	8.5778

domly choose one source, and on its claims about that city, apply Online Truth Prediction algorithm to get the prediction. To evaluate the prediction result, we compare two sets of the predicted truths with the ground truths under MAE and RMSE measures. Table 3 summarizes the results. From the result, we observe that the accuracy of prediction on historical estimated truth is 7.2% better under MAE and 1.6% better under RMSE compared with that of the prediction based on a single source's claims. That means truth discovery component can provide relatively high quality data for time series prediction, which is the reason why truth prediction can benefit from the truth discovery component.

4 Related Work

In this section, we discuss the related work in terms of truth discovery and time series prediction.

Truth Discovery. Motivated by the strong need to resolve conflicts among the noisy data, truth discovery [1, 2, 5, 19, 20, 21, 22, 23] emerges as a hot research topic, due to its ability to estimate source reliability degrees and infer true information from noisy multi-source data. Various truth discovery methods have been proposed to deal with different scenarios, such as complex data types [17, 18], semi-supervised setting [24], source reliability enrichment [3, 4], and output explanation [25]. Nowadays, people have successfully applied truth discovery methods to several applications including but not limited to social sensing [26, 27], knowledge graph [28] and information retrieval [29].

In many real-world applications, data usually come sequentially. To tackle the challenges brought by streaming data, some recent truth discovery algorithms are proposed: Li et. al. [7] and Zhao et. al. [8] present incremental methods that are developed to improve the efficiency of truth discovery. Methods that capture the temporal relations among objects at different timestamps are developed in [6, 30, 31]. However, these methods either ignore the temporal patterns of evolving truths [7, 8] or make the strong smoothness assumption about the objects [6, 30, 31, 32]. Such smoothness assumption may not hold in many real-world applications where time series data with seasonal trends are generated. Garcia-Ulloa [23] proposed a a dynamic graphical model for spatio-temporal event discovery. However, this method cannot deal with data that have seasonal trends. The proposed OTD framework releases the smoothness assumption and can infer the evolving patterns of truths (with/without seasonal trends) to improve the performance of truth discovery. Meanwhile, OTD works in an online fashion and thus has great efficiency when dealing with streaming data.

Time Series Prediction. There are several least square and maximum likelihood based approaches [33, 34, 9, 35] for time series parameter estimation and prediction with independent Gaussian noise assumption. Later, Tsay et. al. [36] develop an iterated least square approach to consistently estimate the parameters of Auto-regressive model, and in [37], least square based and gradient based algorithms are proposed without assuming noise stationarity, ergodicity, or the existence of higher order moments. However, only a few online algorithms are studied: Anava et. al. [12] and Liu et. al. [11] develop online algorithms for ARMA, which uses regret minimization techniques. An online algorithm for time series prediction with the presence of missing data is proposed in [13]. An online adaptive forecasting for time varying auto-regressive processes is developed in [38]. The proposed OTD extends the approach of online ARIMA parameter estimation and prediction to SARIMA model.

As we demonstrated above, the results from time series prediction approaches can help the procedure of truth discovery. Meanwhile, the results of truth discovery can also improve the accuracy of time series prediction. Therefore, by integrating time series prediction with truth discovery, the proposed OTD framework can achieve the best performance on multi-source time series data.

5 Conclusions

In many applications, time series data are continuously generated by multiple sources. In order to extract trustworthy information from the noisy conflicting multi-source data, truth discovery approaches are developed to jointly estimate source reliability and aggregate multi-source data weighted by the estimated reliability. However, existing work on truth discovery fails to capture the temporal patterns embedded in the time series data. Therefore, we propose a novel online truth discovery framework, called OTD, to infer true information from time series data. OTD contains two components: multi-source truth discovery component and time series analysis component. The two components are integrated seamlessly so that they can mutually enhance each other. Through extensive experiments on synthetic and real-world datasets, we demonstrate that the proposed OTD framework can improve the performance of not only truth discovery but also time series analysis.

6 ACKNOWLEDGMENT

This work was supported in part by the US National Science Foundation under grants NSF IIS-1218392, IIS 1514204,

IIS-1319973, CNS-1652503, and CNS-1737590. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] X. Yin, J. Han, *Semi-supervised truth discovery*, in Proc. of WWW, 2011.

[2] X. Li, X. L. Dong, K. B. Lyons, W. Meng, and D. Srivastava, *Truth nding on the deep web: Is the problem solved?* PVLDB, 2012.

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

[4] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, D. Murat, W. Fan, and J. Han, *A condence-aware approach for truth discovery on long-tail data*. PVLDB, 2015.

[5] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han, *A survey on truth discovery*. ACM SIGKDD Explorations Newsletter, 2016.

[6] Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, and J. Han. *On the discovery of evolving truth*. In Proc. of KDD, 2015.

[7] Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, and J. Han. *Conflicts to harmony: A framework for resolving conflicts in heterogeneous data by truth discovery*. IEEE Transactions on Knowledge and Data Engineering, 2016.

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

[9] P. J. Brockwell and R. A. Davis, *Introduction to time series and forecasting*. Springer Science & Business Media, 2006.

[10] A. B. Owen, *A robust hybrid of lasso and ridge regression*. Contemporary Mathematics, 2007.

[11] C. Liu, S. C. Hoi, P. Zhao, and J. Sun, *Online arima algorithms for time series prediction*. In Proc. of AAAI, 2016.

[12] O. Anava, E. Hazan, S. Mannor, and O. Shamir, *Online learning for time series prediction*. In Proc. of COLT, 2013.

[13] O. Anava, E. Hazan, and A. Zeevi, *Online time series prediction with missing data*. In Proc. of ICML,2015.

[14] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientic, 1995.

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

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

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

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

[19] R. Pochampally, A. D. Sarma, X. L. Dong, A. Meliou, and D. Srivastava, *Fusing data with correlations*. In Proc. of SIGMOD, 2014.

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

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

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

[23] D. A. Garcia-Ulloa, L. Xiong, and V. S. Sunderam, *Truth discovery for spatiotemporal events from crowdsourced data*. PVLDB, 2017.

[24] X. Yin and W. Tan, *Semi-supervised truth discovery*. In Proc. of WWW, 2011.

[25] X. L. Dong and D. Srivastava, *Compact explanation of data fusion decisions*. In Proc. of WWW, 2013.

[26] C. C. Aggarwal and T. Abdelzaher, *Social sensing*. In Managing and mining sensor data, pages 237-297. 2013.

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

[28] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy, S. Sun, and W. Zhang, *From data fusion to knowledge fusion*. PVLDB, 2014.

[29] D. Yu, H. Huang, T. Cassidy, H. Ji, C. Wang, S. Zhi, J. Han, C. Voss, and M. Magdon Ismail, *The wisdom of minority: discovering and targeting the right group of workers for crowdsourcing*. In Proc. of COLING, 2014.

[30] C. Meng, W. Jiang, Y. Li, J. Gao, L. Su, H. Ding, and Y. Cheng, *Truth discovery on crowd sensing of correlated entities*. In Proc. of Sensys, 2015.

[31] S. Wang, D. Wang, L. Su, L. Kaplan, and T. Abdelza- her, *Towards cyber-physical systems in social spaces: The data reliability challenge*. In Proc. of RTSS, 2014.

[32] S. Wang, L. Su, S. Li, S. Yao, S. Hu, L. Kaplan, T. Amin, T. Abdelzaher, and W. Hongwei, *Scalable social sensing of interdependent phenomena*. In Proc. of IPSN, 2015.

[33] J. D. Hamilton, *Time series analysis*, volume 2. Princeton university press Princeton, 1994.

[34] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[35] S. C. Hillmer and G. C. Tiao, *An ARIMA-model-based approach to seasonal adjustment*. Journal of the American Statistical Association, 1982.

[36] R. S. Tsay and G. C. Tiao, *Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models*. Journal of the American Statistical As- sociation, 1984.

[37] F. Ding, Y. Shi, and T. Chen, *Performance analysis of estimation algorithms of nonstationary arma processes*. IEEE Transactions on Signal Processing, 2006.

[38] C. Giraud, F. Roue, A. Sanchez-Perez, et al. *Aggregation of predictors for nonstationary sub-linear processes and online adaptive forecasting of time varying autoregressive processes*. The Annals of Statistics, 2015.