Abstract—Recently, privacy preserving issues have been actively studied on the time-series data widely used in variety of applications such as financial, medical, and weather analysis. In this paper, we survey and analyze the recent work of privacy preserving data mining (PPDM) on time-series data. For this, we first investigate what is the privacy in time-series data. We then survey various perturbation techniques on time-series data in the centralized computing environment. We next investigate secure multi-party computation (SMC) and encryption techniques in the distributed computing environment. Social network and cloud computing applications incur a large volume of sensitive (time-series) data, and thus, privacy preserving techniques for exploiting those sensitive data have become much more substantial in many research areas. Our survey results can be used for developing efficient and robust time-series data-based PPDM techniques that can be applied to new computing environments.

Keywords—privacy preserving, time-series data, data perturbation, secure multi-party computation

I. INTRODUCTION

Owing to the rapid increasing the amount of data produced and/or collected through Internet and mobile devices in the recent years, the risk of information disclosure is increasing significantly. Therefore, the growing concern is how to store, manage, and analyze a large volume of data while preserving the privacy. The aim of privacy preserving data mining (PPDM in short) is to extract meaningful knowledge and useful patterns from a large volume of data while preserving their confidentiality. Consequently, PPDM has been actively studied with variety of applications after first proposed in the early 2000s[1-4][10][13].

The real challenge of analyzing the data, concerning privacy, has justified the need of privacy protection in time-series data as well. Time-series data mining, a method of discovering hidden information or knowledge from time-series data, includes pattern discovery, clustering, classification, and rule discovery. In general, original time-series are provided to data miner (mainly 3rd party) in order to perform the mining process. However, the time-series data such as fingerprint, voice, and electrocardiography data are very sensitive since they have confidential information of individuals. Therefore, data providers (owners) do not prefer to provide their original time-series in order to preserve privacy. As shown in Fig. 1, recent techniques to get the same (or similar) mining results preserving privacy at the same time have gained the interest of researchers.

Fig. 1 Privacy preserving time-series data mining.

This paper surveys and analyzes data perturbation techniques and distributed privacy preserving techniques separately for time-series data. Data perturbation techniques publish only the perturbed time-series instead of the original time-series to hide the sensitive information. First, perturbation techniques include noise addition, compression-based perturbation, geometric transformation perturbation, and k-anonymity. Second, privacy preserving techniques in the distributed computing environment, where multiple data providers get common mining results while preserving privacy of their own data, mainly use SMC and encryption techniques.

The rest of this paper is organized as follows. In Section 2, we investigate the sensitive characteristics of time-series data. In Section 3, we survey and analyze the recent work of perturbation techniques on time-series data. In Section 4, we survey the privacy preserving techniques in the distributed computing environment. In Section 5, we finally conclude the paper.

II. PRIVACY OF TIME-SERIES DATA

A time-series is a sequence of real numbers representing values at specific time points (or calculated per fixed interval). Typical examples of time-series data include stock prices, exchange rates, biomedical measurements, weather data, voice data, and fingerprints data[3]. Moreover, time-series data usually have the characteristics of high dimensionality and the
changes in data values over time such as peak, trough, trend, and periodicity. We may need to keep their confidentiality in the data mining process since these characteristics individually can be considered as sensitive information as follows[2][3].

- Amplitude indicates the strength of a signal.
- Representing extreme situations, peak and trough may disclose extreme changes.
- By observing trends of time-series data, an adversary may predict future changes of time-series data.
- Periodicity indicates the periodic changes in time-series data.

Therefore, perturbation techniques to get the meaningful mining results while hiding the sensitive original time-series have been actively studied in recent years[3][7-10]. Time-series data can also be easily reconstructed from the perturbed data because time-series data consists of real numbers. Thus, preserving privacy in time-series data mining process as well as privacy of the time-series data itself is also required in PPDM.

III. TIME-SERIES PERTURBATION

As the most established method to prevent the privacy leakage, data perturbation techniques publish only the perturbed time-series instead of the sensitive original time-series[3][8][9]. In other words, as shown in Fig. 2, multiple independent data sources provide their perturbed time-series only, and a data miner discovers meaningful mining results from the perturbed data. However, the privacy of the original data can be still revealed to third party or attackers by analyzing the mining results, even if the published data are perturbed. In order to prevent this problem, the degree of perturbation should be strong, which may incur the problem of decreasing mining accuracy even though privacy is preserved[3]. In general, privacy preservation and mining accuracy can be controlled by the degree of perturbation, so perturbation techniques are mainly evaluated by these two metrics, namely the degree of the privacy preservation and the loss of the information.

A. Noise Addition

Adding noise is a technique that hides the sensitive data by adding random noise to the original time-series, and it is categorized into simple additive noise (SAN) and multiplicative noise (MN). For an original time-series \( X \) and a random noise time-series \( E \), SAN and MN provide to data miner only \( Y = X + E \) and \( Y = X \times E \), respectively. Fig. 3 shows the example of data perturbation by additive random noise. Fig. 3(c) is a resulting time-series data by adding a noise time-series (Fig. 3(b)) to an original time-series (Fig. 3(a)). In this example, privacy of the original time-series is preserved by providing Fig. 3(c) instead of Fig. 3(a). However, random noise can be removed by the noise filtering attacks because it has a predictable structure. Kargupta et al.[4] have empirically demonstrated that the random noise preserves a very little amount of data privacy since most of noise can be removed if its variance is not large enough.

For the purpose of preventing the filtering attacks, [5] and [6] have exploited data correlations in generating the noise. Huang et al.[5] have used the Principal Component Analysis (PCA) and the Bayes Estimate (BE) for data reconstruction and noise generation based on data correlations. They have pointed out that simple random noise is added evenly in both principal and non-principal components while most of the information of the original data concentrates on principal components, and they have proposed the correlated random noise. These techniques are based on the idea that the random noise becomes difficult to be filtered from the original data if it is concentrated on principal components only. Based on the idea, they guarantee that the noise is concentrated on the principal components by making the correlations of the random noise similar to the correlations of the original data. Mukherjee et al.[6] have proposed the distance-based privacy preserving technique that combines the favorable features of additive perturbation and the orthogonal transformation to avoid the correlation-based and transform-based attacks.

Existing random perturbation techniques are not effectively applicable to preserve distance order preservation among time-series, so this incurs the problem of decreasing the mining accuracy[7]. Moon et al.[3] have proposed a perturbation technique that tries to preserve distance orders. To preserve distance orders among time-series, they use the noise averaging effect of piecewise aggregate approximation (PAA), which is derived from an intuition that the summation of white noise eventually converges to 0 since the mean of noise is 0. They exploit the noise averaging effect using their averages of multiple entries instead of individual entries in computing the distance between distorted time-series, and the effect improves the accuracy of mining results.

Jin et al.[8] have proposed the region-based perturbation technique. They partition all time-series spatially into accord and discord regions by analyzing the impact of local regions on overall classification performance, and they perturb original time-series differentially according to their positions. Thus, they improve the mining accuracy while preserving the local patterns in regions.
B. Compression-based Perturbation

Transformation is a process of converting a high-dimensional time-series into a new feature space of lower dimension, and it is applied to get feature vectors of lower dimension from time-series. A high-dimensional time-series is usually transformed into a few feature vectors to be indexed into multi-dimensional trees such as R-tree for the fast search. Transformational examples include discrete Fourier transform (DFT), discrete Wavelet transform (DWT), and singular value decomposition (SVD). In particular, DFT and DWT are often used in preserving privacy of time-series since they have a property of Euclidean distance preservation [7][9][10].

Mukherjee et al. [7] have used Fourier transformation techniques for the first time in privacy preservation. They preserve the privacy of the original data by using only several Fourier coefficients instead of the original time-series in the mining process. Kim et al. [10] have pointed out that the previous DFT coefficient technique [7] has the critical problem in privacy preservation that the original data might be reconstructed by inverse DFT, and they have proposed Fourier magnitude-based privacy preserving techniques. Their solutions make the reconstruction of original time-series difficult by using only DFT magnitudes except DFT phases.

Recently, Wavelet-based privacy preserving techniques have provided significant importance [9][11]. Papadimitriou et al. [9] have proposed DFT and DWT-based privacy preserving techniques. They observe that the noise is added equally to all coefficients while most energy is concentrated on a few transformation coefficients. Based on this observation, they have perturbed only “important” coefficients whose magnitudes greater than the given threshold.

C. Geometric Transformation Perturbation

Perturbation techniques may change correlations among dimensions as well as primary properties of data influencing in the mining results. Moreover, random data perturbation has the problem of ignoring correlations among dimensions by processing each dimension independently, whereas DFT and DWT transformation-based privacy preserving methods focus on dimensionality reduction instead of considering correlations among dimensions. For the purpose of disguising sensitive information and correlations among dimensions, geometric data transformation techniques have proposed in [20] and [21].

Geometric transformation-based perturbation approaches exploit rotation, translation, and scaling and perturb all dimensions at the same time in order to preserve correlations among dimensions unlike the existing perturbation techniques. Chen et al. [21] have proposed a rotation perturbation technique as follows: for original data set \(X = [x_1, \cdots, x_n]\) and rotation matrix \(R_{d \times d}\) (where \(x_i\) represents a vector), it perturbs through geometric rotation \(g(X) = RX\). To estimate the degree of privacy for a rotation perturbation method, they extend the variance-based privacy metric from a single dimension to a multi-dimension unified metric.

Also, Chen et al. [21] have addressed potential attacks to geometric perturbation, those are, ICA-based attacks, attacks to the rotation center, and distance-inference attacks. Fig. 4 shows the weak points of the basic rotation perturbation. The rotation perturbation may incur the problem that the points around the origin can remain close to the origin even after the perturbation since it uses the origin as the rotation center. Also, it is possible that the original data can be reconstructed by mapping between points and distances, if several points are disclosed. Chen et al. [21] and Mohaisen et al. [20] have proposed some advanced techniques to complement disadvantages of the rotation perturbation.

D. Comparison of Existing Perturbation Techniques

Table I makes a summary of characteristics for existing data perturbation techniques. As shown in the table, existing

<table>
<thead>
<tr>
<th>Existing techniques</th>
<th>Random noise</th>
<th>Correlation-based noise</th>
<th>Dimensionality reduction</th>
<th>Distance preservation</th>
<th>Mining application</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>○</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>[3]</td>
<td>○</td>
<td>X</td>
<td>X</td>
<td>○</td>
<td>Clustering</td>
</tr>
<tr>
<td>[5]</td>
<td>X</td>
<td>○</td>
<td>○</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>[8]</td>
<td>X</td>
<td>○</td>
<td>X</td>
<td>X</td>
<td>Classification</td>
</tr>
<tr>
<td>[7]</td>
<td>X</td>
<td>X</td>
<td>○</td>
<td>○</td>
<td>Classification</td>
</tr>
<tr>
<td>[10]</td>
<td>X</td>
<td>X</td>
<td>○</td>
<td>○</td>
<td>Clustering</td>
</tr>
<tr>
<td>[20]</td>
<td>X</td>
<td>X</td>
<td>○</td>
<td>△</td>
<td>Clustering</td>
</tr>
<tr>
<td>[9]</td>
<td>X</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>[21]</td>
<td>○</td>
<td>X</td>
<td>X</td>
<td>△</td>
<td>Classification</td>
</tr>
</tbody>
</table>

(○: positive, X: negative, △: semi-positive)
techniques are classified into random noise, correlation-based noise, dimensionality reduction, distance preservation, and mining application. [1] and [3] add evenly the degree of noise into all dimensions. On the other hand, [5] and [8] add the noise differently to different dimensions (or coefficients) by using the correlation-based noise. In particular, [3] tries to preserve distance orders among perturbed time-series by exploiting the noise averaging effect. [7] and [10] preserve the privacy of the original data by using only Fourier (or Wavelet) coefficients instead of the original time-series in the mining process; [20] use the rotation in data perturbation. [20] partially preserves scalar products and distances among vectors while mitigating the ICA-based attacks through multiple rotations. It is simple to apply adding noise techniques to existing mining algorithms and possible to control the degree of the privacy preservation and the mining accuracy according to the amount of adding noise. Thus, the adding noise technique is often combined with different privacy preserving techniques. [21] is combined into the adding noise technique with geometric transformation perturbation, and [9] is combined into the additive noise technique with compression-based perturbation.

IV. DISTRIBUTED PRIVACY PRESERVATION

Distributed privacy preservation techniques are applied when the data are distributed in multiple nodes. In such techniques, each data provider directly takes part in the mining process. As shown in Fig. 5, each node performs its own mining process, and then it shares the intermediate mining results with other nodes or transmits to an end node. The end node elicits the final mining results by aggregating the intermediate results of individual node. In the mining process, SMC is usually used to preserve the privacy of time-series. SMC is a technique that computes aggregate values such as sum, average, and distance without disclosing the original time-series, and SMC protocols for several operations have been proposed after A. C. Yao[24] had proposed SMC for comparison. Also, secure solutions using primitive operations of SMC protocols such as clustering, classification and rule discovery have been proposed [13][14][17].

![Fig. 5 Privacy preservation in the distributed environment.](Image)

A. Secure Scalar Product and Secure Euclidean Distance

Several distance functions such as Euclidean and DTW(dynamic time warping) can be used to evaluate the degree of similarity between time-series in clustering, classification, and similarity search. If the scalar product can be computed securely, Euclidean distance and cosine similarity can also be computed securely. Therefore, secure computation on scalar product has been studied in [12-13] since privacy of time-series can be preserved in a variety of applications.

Homomorphic encryption[12], random matrix[13], and secure intersection have been used to compute scalar product of two private vectors while preserving privacy. Du et al.[12] have proposed the homomorphic encryption-based protocol. They also use the data perturbation technique of adding random numbers to the original data to hide the original vectors. Vaidya et al.[13] have tried to preserve the privacy by sending $X' = x_1 + c_{1,1} * R_1 + c_{1,2} * R_2 + \cdots + c_{1,n} * R_n$ generated by $R_1, \cdots, R_n$ instead of sending $X$ in vertically partitioned data. However, Goethals et al.[14] have demonstrated that scalar product protocols of [12] and [13] are not private. Besides, Wong et al.[22] have proposed an asymmetric scalar-product preserving encryption that is not distance-recoverable and focused on the problem of k-nearest neighbor (k-NN) computation over an encrypted database. This technique allows to preserve privacy by encrypting each time-series differently even though query and data time-series are the same.

B. Privacy Preserving Query Processing

In the distributed privacy preserving data mining, query time-series along with data time-series are the major concerns. Processing range query or k-NN query without disclosing any information for both of query and data time-series is the motivational force behind privacy preserving query processing. These techniques process the query directly on an encrypted database or use SMC protocols. Agrawal et al.[23] have proposed an order-preserving encryption scheme for numeric data, and this encryption technique allows to process query directly without any decryption process.

In order to process secure range query on the multi-dimensional data, Chen et al.[16] have proposed the RA(n)dom SP(a)ce Encryption(RASP) approach. RASP has two important features that it does not preserve the ordering of dimensional values, but it is convexity preserving. In addition, to process the query securely as an SMC protocol, Hu et al.[17] have proposed the secure protocol based on the homomorphic encryption which processes k-NN queries on R-trees in the cloud computing environment. Shaneck et al.[18] show how privacy preserving nearest neighbor search can be used in outlier detection, shared nearest neighbor, clustering, and k-NN classification.

C. Privacy Preserving Aggregation

Recently, researches are underway to reduce the risk of privacy leakage incurred by recurring queries of an aggregate query such as sum and count on the distributed time-series[15][19]. Rastogi et al.[19] have proposed the framework called PASTE, which combines Fourier perturbation algorithm(FPAk) and distributed Laplace perturbation algorithm(DLPA).
Shi et al. [15] have proposed private stream aggregation algorithms. They consider a construction framework that multiple participants periodically upload encrypted values to a data aggregator. In particular, they add noise before encrypting the individual time-series. Namely, each participant provides $c_i = E(s_k, X_i + r_i)$ to a data aggregator, where $c_i$ is a value encrypting time-series $X_i$ with noise $r_i$, and $sk_i$ is a private key for each participant. A data aggregator decrypts the noisy sum from multiple ciphertexts and guarantees the differential privacy since decryption results are integrated with noise.

**V. CONCLUSION**

In this paper, we survey and analyze the recent work of privacy preserving data mining on time-series data. Previous PPDM techniques are difficult to be applied to time-series data efficiently because of their unique characteristics. In other words, the previous techniques cannot be directly applied to the metric-based mining since they do not take into account of the high-dimensional characteristic of time-series data, and the perturbed data do not ensure distance among objects. Therefore, the paper surveys data perturbation techniques in the centralized computing environment and the distributed privacy preserving techniques and investigates the characteristics of each of various techniques. As social network and cloud computing applications incur a large volume of sensitive data, privacy preserving techniques for exploiting these sensitive data have become much more substantial for a variety of applications. Our survey results can be used for developing efficient and robust time-series data-based PPDM techniques that can be applied to new computing environments.

**ACKNOWLEDGMENT**

This work was partially supported by Defense Acquisition Program Administration and Agency for Defense Development under the contract.

**REFERENCES**


