VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Anonymization of Time Series Data

J Component

Review – 1

Submitted for

Course: Data Privacy

Course Code: BCI2001

Slot: B2+TB2

Submitted to

Prof. Jasmin T Jose

Team Members

- Keerthivasan K (20BCI0193)
- Pratheeshkumar N (20BCI0195)
- Mukundhan D (20BCI0291)
- Aravinth M (20BCI0192)
- Poovarasan (20BCI0194)

ABSTRACT:

The aim of our project is to anonymize Time Series Data to prevent them against linkage attacks and to preserve the pattern with the help of K-P- Anonymity algorithm. Some libraries that we used in implementation are NumPy, Pandas,Loguru and Saxpy.

A sequence of observations indexed by the time of each observation is called a time series.

Time Series data have a very complex structure. They are used for various purposes such as forecasting or prediction study of underlying processes in healthcare pattern discovery and so on Therefore, when transforming/anonymizing time series data, the anonymized data should be useful and provide accurate results in these applications.

The data set contains three disjoint sets of data. Explicit Identifiers (EI) such as SSN and names. Quasi- identifiers (QIs) contain a series of time related data (A1,…, AN). Sensitive attributes are a series of time-related data that are considered sensitive and should not be altered.

They first study achievability results for the case where the time-series of users are governed by an i.i.d. process. The converse results are proved both for the i.i.d. case as well as the more general Markov chain model.

Literature Review

Sl. No.	Year	Authors, Title	Contributions	Limitations
1	2011	Lidan Shou, Xuan Shang, Ke Chen, Gang Chen, and Chao Zhang, Supporting Pattern-Preserving Anonymization for Time-Series Data	They studied the anonymization of time series and said why the conventional k-anonymity model cannot effectively address this problem as it may suffer severe pattern loss. Proposed a novel anonymization model for pattern-rich time series. This model publishes both the attribute values and the patterns of time series in separate data forms.	The current solution imposes a very strict constraint on PR equality and this may cause serious pattern loss.
2	2021	Jian Xu1 Wei Wang1 Jian Pei2 Xiaoyuan Wang1 Baile Shi1 Ada Wai-Chee Fu, Utility-Based Anonymization for Privacy Preservation with Less Information Loss	They have studied the problem of utility-based anonymization. A simple framework was given to specify utility of attributes, and two simple yet efficient heuristic local recoding methods for utility-based anonymization were developed. The bottom-up method and the top-down method achieve better anonymization than the MultiDim.	The computation time is often a secondary consideration yielding to the quality.
3	2017	Vincent thouvenot, damien Nogues, Catherine Gouttas, Data-driven anonymization process applied to time series	Digital transformation and Big Data allow the use of highly valuable data. However, these data can be individual or sensitive, and represent an obvious threat for privacy. Anonymization, which achieves a trade-off between data protection and data utility, can be used in this context.	It describes a data-driven anonymization process and apply it on simulated electrical load data
4	2015	Dymitr Ruta, Ling Cen, Ernesto Damiani, Fast summarization and anonymization of multivariate big time series	Data anonymization is expected to solve this problem, yet the current approaches are limited predominantly to univariate time series generalized by aggregation or clustering to eliminate identifiable uniqueness of individual	Implementation of the anonymizing summarization involves shape preserving greedy elimination and aggregation that supports parallel cluster processing for big data implementation

			data points or patterns. For multivariate time series, uniqueness among of the combination of values or patterns across multiple dimensions is much harder to eliminate due the to exponentially growing number of unique configurations of point values across multiple dimension	
5	2014	Stephan Kessler, Erik Buchmann, Thorben Burghardt, Klemens Bo˙hm, Pattern-sensitive Time-series Anonymization and its Application to Energy-Consumption Data	Time series anonymization is an important problem. One prominent example of time series are energy consumption records, which might reveal details of the daily routine of a household. Existing privacy approaches for time series, assume that every single value of a time series contains sensitive information and reduce the data quality very much	They Proposed (n,l,k) Anonymity To reduce the Information loss , But it does not work properly for univariate series
6	2013	Lidan Shou,Xuan Shang,Ke Chen,Gang Chen,Chao Zhang, Supporting Pattern-Preserving Anonymization for Time-Series Data	Time series is an important form of data available in numerous applications and often contains vast amount of personal privacy. The need to protect privacy in time-series data while effectively supporting complex queries on them poses nontrivial challenges to the database community. We study the anonymization of time series while trying to support complex queries, such as range and pattern matching queries, on the published data.	This model publishes both the attribute values and the patterns of time series in separate data forms.
7	2020	J.S.Adeline Johnsana, A.Rajesh, S.Sangeetha and S.Kishore Verma, Value and Pattern Anonymization of Time Series Data for Privacy Preserving Data Mining	To protect privacy on time series data more number of techniques have been proposed, out of which the conventional k-anonymity picked up the significance, yet it comes up short in giving limited protection to the patterns of the time series data as it might endure extreme pattern loss	In this paper a combination of novel methodologies K-anonymization (SKY), Symbolic polynomial with cross validation for pattern representation is proposed to reveal a promising level information loss and pattern loss for the

				Privacy Preserved Data Mining field of exploration
8	2016	Mohammad - Reja Pajoohan, Privacy Preservation for Publishing Medical Time Series: *k*- anonymization of Ngram	In this paper, they address this problem and define the k- anonymity principle for the Ngram. The proposed schema aims to provide the k-anonymization by repeating the rare n-grams to hide them in the crowd of frequent n-grams.	This Algorithm provides low entropy information loss
9	2019	Erik Wik, Anonymously Publishing Univariate Time-Series	In this thesis, an investigation is made into how to protect univariate time-series. The main focus is on publish- ing anonymized time-series from individual users, but methods for anonymizing aggregate time-series and the removal of sensitive data is also investigated. This is done in order to find a wider understanding of how a blood glucose related database can be anonymized	The results show that PC-KAPRA offers a large improvement in retaining pattern information compared to KAPRA, and publishes data which could be considered qualitative useful information
10	2017	Nazanin Takbiri , Amir Houmansadr , Dennis Goeckel, Hossein Pishro-Nik, Matching Anonymized and Obfuscated Time Series to Users' Profiles	Many popular applications use traces of user data to offer various services to their users. However, even if user data is anonymized and obfuscated, a user's privacy can be compromised through the use of statistical matching techniques that match a user trace to prior user behavior. In this research paper, they derive the theoretical bounds on the privacy of users in such a scenario.	They first study achievability results for the case where the time-series of users are governed by an i.i.d. process. The converse results are proved both for the i.i.d. case as well as the more general Markov chain mode
11	2017	U.S. Census records With Demographic attributes	Population based risk • Dataset based risk • Safe harbor	Complexity issue • Individual record can be re-identified
12	1994	*US Census (USC)*, an excerpt of records from the 1994 U.S. Census database which is often used for evaluating anonymization algorithms	In this research, authors have presented a data publishing algorithm that satisfies the differential privacy model. • One of the features which the research has is: The	Results are not accurate in case of both i.e. categorical and numerical data. Information loss is more.

			transformations performed are truthful i.e. the dataset does not use any input or out perturbation of external data. Records are randomly selected from the given dataset which ensures that the unique feature of certain biomedical aspect remains hidden.	

Gaps Identified

Based upon the several research papers on privacy preserving data publishing and data generalization techniques, the following research gaps can be formulated.

A. Perturbation of Time Series Data with White Noise

In this approach, white noise that is at high frequency is added to time series data, which results in perturbation of values in the original time series data. This approach protects the data by perturbing the values of the original time series data. The utility of the anonymized data set is better when compared with other methods.

Transformed data retain most of the statistical properties of the original time series data set: Preserve the pattern, retain frequency-domain properties, and so on. But the drawback is that it has poor privacy level.

B. Perturbation of Time Series Data with Correlated Noise

Perturbation with correlated noise changes the values of time series data: the pattern and the frequency. This affects the utility of data but provides higher privacy.

Re-identification of time series data perturbed by correlated noise is possible with a regression model.

An adversary can use his background knowledge to implement linear regression model to protect the values.

K-Anonymity

K-anonymity is a key concept that was introduced to address the risk of re-identification of anonymized data through linkage to other datasets. For k-anonymity to be achieved, there need to be at least k individuals in the dataset who share the set of attributes that might become identifying for each individual.

K-anonymity might be described as a ‘hiding in the crowd’ guarantee: if each individual is part of a larger group, then any of the records in this group could correspond to a single person. K-Anonymization is used to prevent linkage attacks, where QI attributes in a record are generalized to be identical with k-1 records. However, the issue with this approach is that with higher levels of generalization, the pattern of the anonymized data set could get distorted.

User Based Categorization of Data:

The existing data publishing algorithms focuses upon different categorization of the data based upon different level of generalization. The parameters like high, medium, low level generalization have been performed and tested to ensure that different kind of data can be visible to preserve its privacy. The same problem can be looked upon on different view where user can be categorized based upon it's role and authenticity viz. role based access model (RABC). If the system allows us to check the credibility and authenticity of the data as well as user before data publishing, then the appropriateness of the PPDP can be maintained. However, such work can go into the category of empirical kind of research where the new parameters are required to test the system. The parameters like loss, entropy etc. may not be sufficient to test the validity of the system.

Uniform Model for PPDP with Role based access model:

From the available literature review, there is no uniform model which incorporates the level of data generalization and role of user simultaneously. Several data publishing techniques which have been studied are only based upon the type of data which is there in the healthcare repository.

Re-identification Attacks Countermeasures

Most of the data publishing techniques have still a good probability of re-identifying the particular tuple from healthcare dataset. No research fruitfully guarantees the re-identification attack will happen. Few of the research papers where privacy achieved is very high, has lots of information loss.

Maintaining variable privacy utility threshold depending upon the priority of health-care data.

Privacy-Utility is always a concern in data publishing. Several latest literature surveys which are cited in this report have agreed on the fact that – both privacy and utility cannot be achieved with highest threshold. There is certain research where privacy parameters have outperformed with maximum information loss and vice versa. Maintaining trade-of between privacy and utility is the major challenge from available literature. Computational complexity of data publishing algorithms. There are several literature based upon the data publishing strategy have more computational complexity (CPU Cycles, Generalization time etc.). Some algorithms outperforms better only if the size of dataset is small. Some algorithms are fruitful only for low or medium dimensional data. There are no fruitful schemes where multi-dimensional sparse data.

Problem Statement

From a privacy perspective, this implies that demographics which are a form of QIDs that can be used to reidentify an individual when combined with other publicly available data can be safely removed without affecting the prediction result.

Similarly, the remaining (non-private) data can be released without violating patient's privacy. In this paper, we show that this approach of releasing data still poses a threat to privacy. Assume that an adversary has some demographic information of some data records, she can train a classifier or ML model using the released data (with the demographics data removed) and the subset of demographic information that she has to predict the demographics of the entire dataset. That is, the demographic information of the rest of the patients, for whom she does not have can be inferred. This implies that demographics encodes some amount of information and in the presence of other attributes, demographics seems not to be useful (i.e. it is correlated with other attributes).

Reference

1. Shou, L., Shang, X., Chen, K., Chen, G., & Zhang, C. (2011). Supporting pattern-preserving anonymization for time-series data. IEEE Transactions on Knowledge and Data Engineering, 25(4), 877-892.
2. Zheng, W., Wang, Z., Lv, T., Ma, Y., & Jia, C. (2018, November). K-anonymity algorithm based on improved clustering. In International conference on algorithms and architectures for parallel processing (pp. 462-476). Springer, Cham.
3. Yin, C., Zhang, S., Xi, J., & Wang, J. (2017). An improved anonymity model for big data security based on clustering algorithm. Concurrency and Computation: Practice and Experience, 29(7), e3902.
4. T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein, Introduction to Algorithms, second ed. MIT press and McGraw-Hill, 2018.
5. R. Dewri, I. Ray, and D. Whitley, "On the Optimal Selection of k in the k-Anonymity Problem," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 1364-1366, 2018.
6. D. Gunopulos and G. Das, "Time Series Similarity Measures," Proc. Tutorial Notes of the Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (Tutorial PM-2), pp. 243-307, 2010.
7. E. Keogh and T. Folias, "UCR Time Series Data Mining Archive," http://www.cs.ucr.edu/ eamonn/TSDMA/, 2012.
8. E.J. Keogh, K. Chakrabarti, S. Mehrotra, and M.J. Pazzani, "Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases," Proc. ACM SIGMOD Conf., pp. 151-162, 2011.
9. E.J. Keogh, K. Chakrabarti, M.J. Pazzani, and S. Mehrotra, "Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases," Knowledge Information Systems, vol. 3, no. 3, pp. 263-286, 2011.
10. E.J. Keogh and M.J. Pazzani, "An Enhanced Representation of Time Series which Allows Fast and Accurate Classification, Clustering and Relevance Feedback," Proc. Fourth Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 239-243, 2008.