DATA PRIVACY FINAL REVIEW

Anonymization of Time Series Data

Aravinth M	- 20BCI0192
Keerthivasan K	- 20BCI0193
Poovarsan A	- 20BCI0194
Pratheeshkumar N	- 20BCI0195
Mukundhan D	- 20BCI0291

Outline

Introduction

Key Challenges

Abstract

Literature Survey

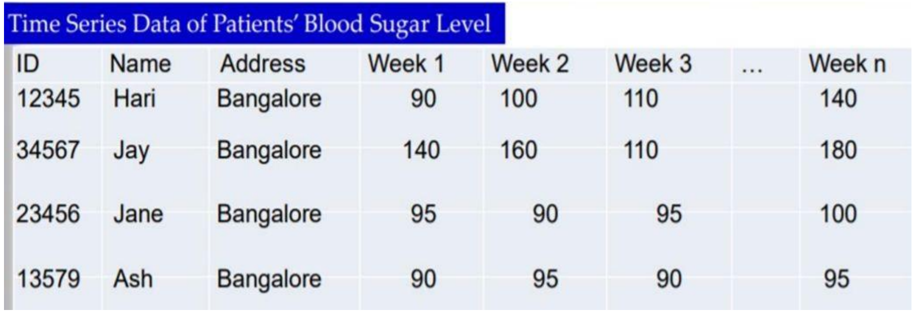
Existing Methods

Proposed Model

References

Introduction

A sequence of observation indexed by the time of each observation is called a time series. Time series data are used for various purposes such as forecasting or prediction study of underlying processes in healthcare pattern discovery and so on. Privacy protection in the publication of time series is a challenging topic mostly due to the complex nature of the data and the way that they are used.



Time Series Data of Patients' Blood Sugar Level

ID	Name	Address	Week 1	Week 2	Week 3	...	Week n
12345	Hari	Bangalore	90	100	110		140
34567	Jay	Bangalore	140	160	110		180
23456	Jane	Bangalore	95	90	95		100
13579	Ash	Bangalore	90	95	90		95

A time series 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

Key Challenges

- **High Dimensionality:** Univariate time series data of 500 values have 500 dimensions to choose from. Protecting high-dimensional data is a problem that does not have an effective solution. Moreover, high-dimensional data coupled with the unknown background knowledge of the adversary make their privacy protection a major challenge.
- **Background Knowledge of Adversary:** Modeling background knowledge of the adversary is not possible. Because of this, the data protection method may lead to high protection or low protection, thus resulting in poor utility.
- **Pattern Preservation:** Time series data have both instant values and a pattern. Any privacy preserving method should ensure that the patterns in the anonymized data set should be preserved as much as possible.
- **Preservation of Statistical Properties:** Time series data exhibit certain statistical properties such as mean, variance, and so on. Any privacy preservation method should ensure that these properties are maintained in the anonymized data set.

Abstract

In this project, we propose **KP-Anonymity algorithm** to anonymize Time Series Data to prevent them against linkage attacks and to preserve the pattern.

Literature Survey

Sno	Name	Year	Journal	Authors	Technique/ algorithm used	Limitations
1	Supporting Pattern-Preserving Anonymization for Time-Series Data	2011	IEEE	Lidan Shou, Xuan Shang, Ke Chen, Gang Chen, and Chao Zhang	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	Utility-Based Anonymization for Privacy Preservation with Less Information Loss	2021	ACM	Jian Xu1 Wei Wang1 Jian Pei2 Xiaoyuan Wang1 Baile Shi1 Ada Wai-Chee Fu	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	Data-driven anonymization process applied to time series	2017	IEEE	Vincent thouvenot, damien Nogues, Catherine Gouttas	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

Sno	Name	Year	Journal	Authors	Technique/ algorithm used	Limitations
4	Fast summarization and anonymization of multivariate big time series	2015	IEEE	Dymitr Ruta, Ling Cen, Ernesto Damiani	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 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	implementation of the anonymizing summarization involves shape preserving greedy elimination and aggregation that supports parallel cluster processing for big data implementation.
5	Pattern-sensitive Time-series Anonymization and its Application to Energy-Consumption Data	2014	IEEE	Stephan Kessler, Erik Buchmann, Thorben Burghardt, Klemens Bohm	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, e.g., from the field of trajectory anonymization, 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

Sno	Name	Year	Journal	Authors	Technique/ algorithm used	Limitations
6	Supporting Pattern-Preserving Anonymization for Time-Series Data	2013	IEEE	Lidan Shou,Xuan Shang,Ke Chen,Gang Chen,Chao Zhang	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	Value and Pattern Anonymization of Time Series Data for Privacy Preserving Data Mining	2020	Open Journal	J.S.Adeline Johnsana, A.Rajesh, S.Sangeetha and S.Kishore Verma	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	Privacy Preservation for Publishing Medical Time Series: *k*-anonymization of Ngram	2016	Open Journal	Mohammad - Reja Pajoohan	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

Sno	Name	Year	Journal	Authors	Technique/ algorithm used	Limitations
9	Anonymously Publishing Univariate Time-Series	2019	IEEE	Erik Wik	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	Matching Anonymized and Obfuscated Time Series to Users' Profiles	2017	IEEE	Nazanin Takbiri , Amir Houmansadr , Dennis Goeckel, Hossein Pishro-Nik	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

Existing Anonymization Methods for Time Series Data and their drawbacks

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.

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.

Proposed Model: KP-Anonymity

(k,P) anonymity is a new model proposed in which P is a new privacy constraint which acts against linkage attacks since pattern preservation is very important in time series anonymization.

This model helps in publishing both attribute values and patterns of time series in separate data forms which ultimately prevents pattern loss. Also, this model supports wide range of queries on anonymized data.

Two levels of anonymization: k, P

On the first level, k-anonymity is required for time series in the entire database. That means the records in the published database can be grouped by the quasi-identifier attribute values, and each group should contain at least k records.

On the second level, P-anonymity is required for the pattern representations associated with each record in a same group. Specifically, each group can be divided into subgroups, each of which contains at least P records having identical pattern representations

Two levels of anonymization: k, P

k-requirement: each anonymization envelope (AE) appears at least k times

P-requirement: for each k-group G, for each time series r in G, there are at least P-1 other time series in G having the same QI pattern representation (PR[r])

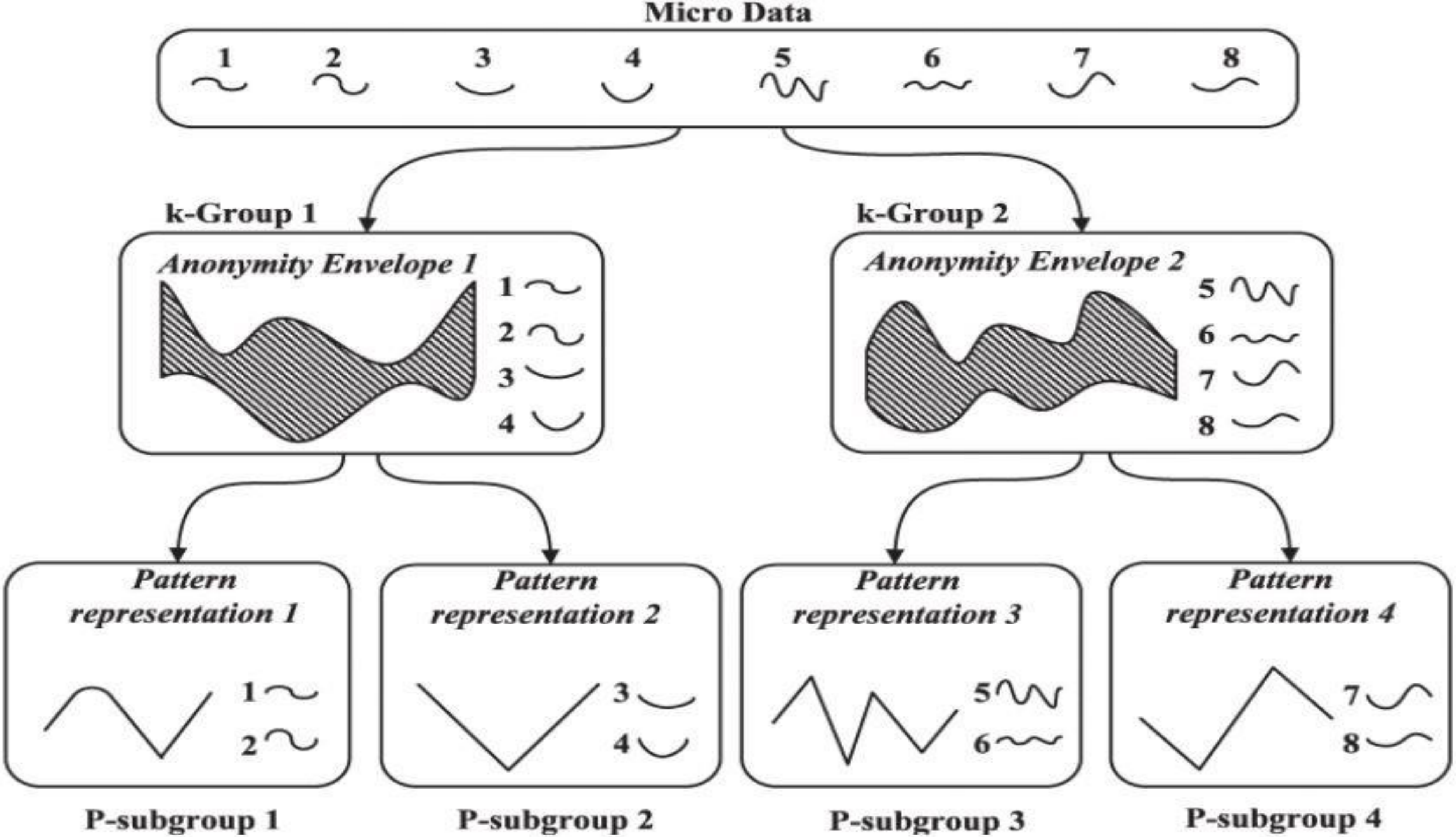
Symbolic Aggregate approXimation (SAX)

k-anonymity →

- Conventional solution for prevent linkage attacks.
- Preserve statistics of the original data.
- Can’t effectively preserve patterns

(k,P)-anonymity →

- Prevent linkage and pattern disclosure attacks,
- Ensure anonymity on two levels: k-anonimity and P-anonymity
- Preserve patterns of time series


Micro Data
1
2
3
4
5
6
7
8
k-Group 1
k-Group 2
Anonymity Envelope 1
1
2
3
4
Anonymity Envelope 2
5
6
7
8
Pattern representation 1
1
2
Pattern representation 2
3
4
Pattern representation 3
5
6
Pattern representation 4
7
8
P-subgroup 1
P-subgroup 2
P-subgroup 3
P-subgroup 4

Conclusion

We proposed a novel anonymity model called (k, P) anonymity for time-series data. Relying on a generic definition to pattern representations, our model could prevent three types of linkage attacks and effectively support the most widely used queries on the anonymized data. Our approach allowed for customized data publishing and provided estimation methods to support queries on such data. The extensive experiments demonstrated the effectiveness of (k, P)-anonymity in resisting linkage attacks while preserving the pattern information of time series. Our results also illustrated the effectiveness and efficiency of the proposed estimation methods for customized data publishing. Our current solution imposes a very strict constraint on PR equality and this may cause serious pattern loss. In the future work, we will consider losing the PR equality condition on the premise of ensuring privacy preservation ability. This strategy may greatly reduce the information loss.

REFERENCES

[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. InInternational 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 clusteringalgorithm. 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 andMcGraw-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.