CSE 1901 – TECHINCAL ANSWERS FOR REAL WORLD PROBLEMS

PROJECT TITLE:

ANONYMIZATION OF USER’S TIME SERIES DATA

PROJECT BASED COMPONENT REPORT

By

Aravinth M – 20BCI0192 (Ph. No: 77080 70428)
Lokeshwaran M – 20BCE2599 (Ph. No: 93612 00265)
Dhamodara prasad S – 20BCE0213 (Ph. No: 98429 56552)

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Anonymization of User’s Time Series Data

20BCI0192 – Aravinth M, 20BCE2599 – Lokeshwaran M, 20BCE0213 - Dhamodara Prasad S

ABSTRACT

Time series data, consisting of sequences of observations indexed by time, are widely used for various purposes, such as forecasting, pattern discovery, and healthcare studies. Toprotect individual privacy and prevent linkage attacks, anonymization of time series data is essential. The dataset contains explicit identifiers (EI), quasi-identifiers (QIs), and sensitive attributes. They are considered for removal during the anonymization process to protect privacy. Since this is a research project which is an iterative process, considering the complexity and depth of this work, it could take anywhere between a few weeks to a month to research and implement the algorithms. Additionally, unforeseen challenges or the need for additional research might extend the timeline further. This research project aims to achieve anonymization while preserving data patterns. The findings highlight the need for a comprehensive approach to privacy protection beyond simple attribute removal. We introduced a new anonymity model, (k, P) anonymity, tailored for time-series data. This model effectively countered various linkage attacks and supported common queries on the anonymized data, without losing the pattern. Our method allowed personalized data publishing and included estimation techniques for query support. Through extensive experiments, we demonstrated the model's efficacy in resisting linkage attacks while retaining time-series pattern information. Our estimation methods also proved efficient for customized data publishing. This technique balances the privacy preservation and patternloss. This approach holds the potential to significantly reduce information loss.

KEYWORDS

Anonymization, Data Privacy, Time series data, KP Anonymization, Data Protection, Data Anonymization.

1. INTRODUCTION

The world is getting smaller, more connected, and more volatile. In this emerging modern world, Data is everything. 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.

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. Because of the complexity of time series data structure, anonymization is rather challenging as there are too many aspects to be taken care of.

This Time Series data contains user data which is very sensitive. This makes this data prone to the attacks made by the attackers/hackers. Therefore, anonymizing and protecting this time series data is very essential to keep the data protected. Therefore, when transforming or anonymizing time series data, the anonymized data should be protected, useful and provide accurate results in these applications.

The main aim of this work would be proposing an algorithm for anonymizingthe time series data. The proposed algorithm should not be prone to attacks such as homogeneous attack, linkage attack, and background knowledge attack on the user time series data.

The algorithm should overcome the key challenges such as high dimensionality, preserve the pattern of the user's time series data, and preserve the usage (utility) of the user's time series data bymaintaining the statistical properties of data.

The objective of the proposed work would be introducing an algorithm called *k-p anonymization* which gives the maximum possible usage (utility) of the user's time series data. This anonymization technique would give us not only maximum possible usage (utility), but also the privacy of the user's time series data.

Time series data is becoming increasingly common and valuable, but it also raises privacy concerns. One way to protect privacy is to anonymize the data, but this can make it difficult to perform queries on the data. The authors of this paper propose a new method for anonymizing time series data that preserves the ability to perform queries. Their method is based on a combination of k-anonymity and a technique called pattern distortion. The authors evaluate their method on a real-world dataset and show that it can effectively preserve privacy while still allowing users to perform queries on the data. Here are some of the key points from the passage: The authors focus on the problem of anonymizing time series data while still allowing users to perform queries on the data. They propose a new method that combines k-anonymity with pattern distortion.

Their method is evaluated on a real-world dataset and shown to be effective in preserving privacy while still allowing users to perform queries.

2. LITERATURE SURVEY

[1] The research paper focuses on addressing the privacy concerns related to user data collected by User-Data Driven (UDD) services, where user data is used to offer various services. The paper considers scenarios where user data is anonymized and obfuscated, but still vulnerable to statistical matching techniques that can compromise user privacy. The goal is to derive theoretical bounds on user privacy in such scenarios and understand the trade-off between privacy protection and user utility, [2] This paper has 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, [3] The research paper focuses on addressing the privacy concerns related to user data collected by User-Data Driven (UDD) services, where user data is used to offer various services. The paper considers scenarios where user data is anonymized and obfuscated, but still vulnerable to statistical matching techniques that can compromise user privacy. The goal is to derive theoretical bounds on user privacy in such scenarios and understand the trade-off between privacy protection and user utility, [4] The research paper is a comprehensive survey that addresses anonymization techniques for privacy-preserving data publishing (PPDP) across different types of data, including both tabular (relational) and graph (structural) data. The goal is to present a deep understanding of the privacy preservation problem, existing anonymization techniques, challenges, and research directions. The survey systematically categorizes the anonymization techniques into two main categories: relational (tabular) anonymization and structural (graph) anonymization, [5] The research paper presents a novel anonymization method called "Chronos" specifically designed for time series data, and it applies this method to electrocardiogram (ECG) time series data. The goal of the method is to provide privacy protection to the data while retaining the utility of the original data. The method is evaluated on ECG datasets from the MIT-BIH Arrhythmia Database and the Physikalisch Technische Bundesanstalt (PTB) Diagnostic ECG], [6] They used a technique called differential privacy to anonymize the IoT data. Differential privacy adds noise to the data in a way that preserves the overall distribution of the data, while making it difficult for the adversary to learn anything about any individual user's behavior, [7] 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, yetit comes up short in giving limited protection to the patterns of the time seriesdata as it might endure extreme pattern loss , [8] The authors used a technique called probabilistic data swapping (PDS) to anonymize user data traces. PDS is a privacy-preserving mechanism that shuffles the data points of different users in a way that preserves the overall distribution of the data, while making it difficult for an adversary to match the data points to their original users, [9] In this thesis, an investigation is made into how to protect univariate time -series. The main focus is on publishing 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, [10] In this thesis, an investigation is made into how to protect univariate time -series. The main focus is on publishing 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. [11] The research paper evaluates the effectiveness of ARX Data Anonymization and Amnesia using the OSSpal methodology and a public dataset of vaccine-related tweets, concluding that ARX Data Anonymization is recommended due to its betterperformance, though the study's limited tool selection and dataset scope should be considered. [12] This paper establishes theoretical privacy boundaries for user data even after anonymization and obfuscation, considering statistical matching techniques. It extends prior workin location privacy by defining these limits for user time-series data, exploring the privacy-utility tradeoff as user network size increases, delineating regions where all users achieve perfect privacy or none retain privacy. [13] This paper addresses the challenge of web traffic de- anonymization caused by growing encrypted traffic. Unlike conventional methods that analyze entire packets, this study introduces a shallow, temporal analysis of web traffic data packets to identify users based on navigation patterns, without accessing TCP packet content. The approach's efficacy is demonstrated through a performance comparison with a traditional feed-forward neural network, highlighting the significance of temporal data in this context. [14] This paper defines and establishes conditions for perfect location privacy, focusing on anonymization. The study proves that the prior sufficient bounds are tight, revealing the thresholdfor achieving perfect privacy using anonymization, both in an independent and identically distributed (IID) model and Markov chain movement modeling. The research highlights the adversary's ability to recover user locations with high probability based on the number of anonymous observations collected. [15] This work addresses time series anonymization, focusing on cases like energy consumption records that can expose household routines. Unlike methods that compromise data quality, our approach targets combinations of tuples in time series, introducing (n; l; k)-anonymity to preserve privacy with minimal information loss, assuming adversaries may access a few data points. We propose heuristics for achieving (n; l; k)-anonymity and validate our method using syntheticand real data, demonstrating that moderate modifications to time series are effective in meeting privacy requirements.

3. MOTIVATION

Protecting the privacy of high-dimensional time series data while preserving their intricate patterns remains a critical challenge. Existing anonymization methods, such as perturbation with noise and k-anonymity, often compromise either privacy or data utility. In this context, we introduce a novel (k, P) anonymity model. By combining k-anonymity with the new privacy constraint P, our model aims to prevent linkage attacks and enhance privacy. Moreover, we propose adistinctive approach by segregating the disclosure of attribute values and patterns, thus safeguarding pattern preservation without sacrificing data utility. Our study provides two algorithms that enable effective implementation of the (k, P) anonymity model, offering a tailored solution for robust privacy and pattern maintenance in high-dimensional time series data sharing.

4. METHODOLOGY

Time series data is a specific type of data that represents observations or measurements collected or recorded over a continuous period of time. Time series data is taken by noting the values for a particular attribute on different intervals.In a ‘time series’ dataset, each data point isassociated with a specific timestamp or time period, making it a valuable resource for analyzing and understanding trends, patterns, and behaviours that evolve over time. The data is a value of a particular attribute, noted frequently. Time series data is prevalent in various domains and is used for a wide range of applications, including economics, finance, weather forecasting, healthcare, engineering, and more. The raw time series data if not anonymized, is prone to various cyber- attacks. So, the raw time series data is anonymized and the anonymized data is then used for various purposes.

4.1. Dataset Discussion

For the purpose of this research endeavor, the focus is directed towards a crucial facet of modern data science – the anonymization of time series data. The datasets selected for this project encompass daily climate time series data, organized on a year-wise basis. These climate time series datasets are of particular significance as they present an opportunity to address the critical issue of data privacy while preserving the integrity and utility of the information contained therein. The core model for this research project has been methodically developed by utilizing data extracted from the years 2013 to 2017 within the previously mentioned dataset.

In addition to the climate data, another pivotal component of this research revolves around the utilization of Sales Transaction time series Weekly Dataset. Sales data, presented in a time series format at a weekly granularity, representsa critical asset for businesses and analysts seeking to understand consumer behavior, market trends, and economic fluctuations. This dataset captures the intricate dynamics of sales transactions, providing a wealth of insights into buying patterns and seasonal variations. By incorporating this Sales Transaction time series Weekly Dataset into our study, we aim to broaden the scope of our research, exploring theapplication of data anonymization techniques in a context that has significant implications for businesses and economic forecasting.

Through the careful selection and consideration of these datasets, thisresearch aims to contribute to the broader discourse surrounding data privacy and the responsible use of valuable time series data resources.

4.2. K-anonymity algorithm

K-anonymity is a key concept that was already introduced to address the risk of reidentification of anonymized datathrough linkage to other datasets. For k- anonymity to be achieved, there need to beat 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 everyone is part of a larger group, then anyof 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.

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 as modelling background knowledge of the adversary is not possible. Because of this, the data protection method may lead tohigh protection or low protection, thus resulting in poor utility (usage).

4.3. K-p anonymity algorithm

In this section, we introduce the k-p anonymity algorithm to address theproblem described before. The perspective of this algorithm would be improving the existing algorithmic techniques by introducing new algorithms and methods. We also look at the characteristics of the proposed model. In the end of this section, we propose a general framework to publish data conforming to k-p anonymity.

4.4. Characteristics of k-p anonymity

K-p anonymity is an algorithm proposed in which P is a new privacy constraint which acts against linkage attacks since pattern preservation is very important in time series data anonymization.

- This algorithm helps in publishing both attribute values and patternsof time series in separate data forms which ultimately prevents pattern loss.
- Also, this algorithm supports a wide range of queries on anonymized data.
- The algorithm is developed to provide high security, privacy, and utility (usage) to the user’s time series data by anonymizing the time series data.
- This algorithm also reduces the high dimensionality of the time anonymized series data.
- This algorithm helps theanonymized data to be very useful to various tasks such as analysis, survey, senses, etc., by improving the utility (usability) of the data.

4.5 Approaches

For implementing k-p anonymity algorithm on time-series data, we have two approaches.

i. Top-down approach (Naïve method)

ii. Bottom-up approach (Kapra method)

4.6. Implementation

To implement this algorithm in the anonymization tool, instead of using the already existing k-anonymity algorithm, the we have to use the k-p anonymity algorithm which is proposed here. To implement k-p anonymity algorithm in the softwares, the modifications on the code should be performed. The new variable should be introduced. That P variable should preserve the pattern of the time series data according to the privacy and utility requirements.

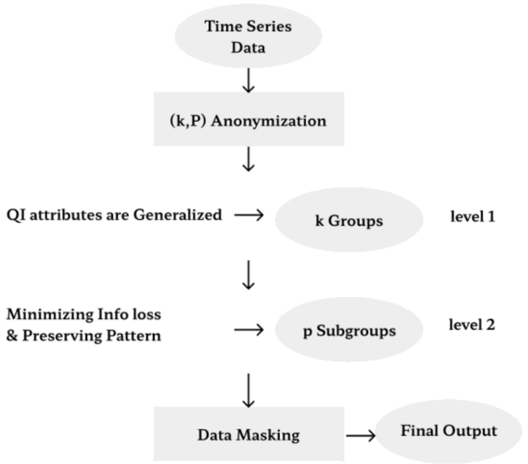


Fig. 1: Work flow Diagram – (k,p) anonymity levels

This anonymity algorithm supports customized data publishing i.e., a certain part of the values and different parts of the pattern of the anonymized time series will be published simultaneously.

It has two phases. Firstly, it performs top-down clustering to ensure k – anonymity of the data set. An additional create-tree procedure is performed for each of the k-groups formed in the first phase.

Our approach assumes that each time series is published in three components, namely the QI value ranges, the QI pattern representation, and the sensitive information. For clarity of presentation, the k-p anonymity algorithm can be described as a conceptual extension of the conventional k-anonymity. Nevertheless, the algorithm to enforce k-p anonymity does not have to rely on the conventional k-anonymity algorithm.

Our algorithm ensures anonymity on two levels.

On the first level, the QI attributes are generalized to fulfil the conventional k-anonymity, regardless of the QI pattern representation. The results of the generalization contain a number of partitions known as the k groups. We note that the QI value ranges are analogous to those in conventional k- anonymity.

The second-level anonymity considers records in each k-group. For any record r in a k- group, if there exist at least P - 1 other records which have the same pattern representation as r, we say that P-anonymity is enforced for this k-group. As a result, we can partition the k-group further into subgroups, each of which contains at least P records having the identical PR. Now, we will look at the method to enforce (k, P)- anonymity on an arbitrary micro data set. Our target is to minimize the information loss while respecting the constraints on the breach probabilities. It can be proven that a globaloptimal solution requires combinatorial computation cost. Therefore, we will consider more efficient near-optimal solutions in the sequel.

i. Top-down approach

Motivated by the conventional k-anonymity, one possible solution for enforcing k-p anonymity is to employ a top-down clustering-like framework as described in the following:

- Generate first-level k-groups from the micro data set.
- For each k-group, extract PRs from micro data based on the chosen PR form. The extracted PRs should minimize the pattern loss while respecting the Pre-requirement within its own k-group.
- For each k-group, generate P-subgroups based on the PRs.

Step 2 is a challenging task and highly dependent on the PR form being used. Different PR forms may lead to very different implementations of this step. In whatever forms, the granularity of PR

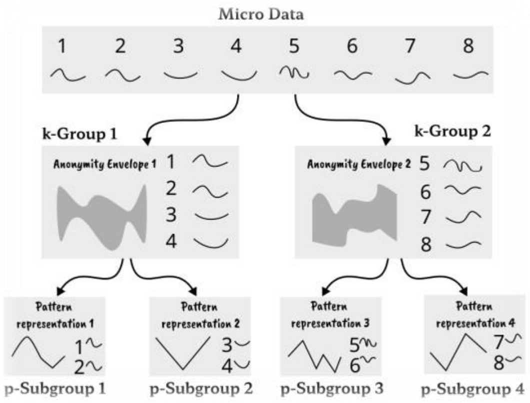


Fig. 2: Framework Diagram – (k,p) anonymity

The top-down approach is easy to understand as it can be regarded as an extension to the existing k- anonymity approach. Alternatively, we can employ a bottom-up framework to form P-subgroups from individual records first, and then build k-groups.

ii. Bottom-up approach

The bottom-up approach is described inthe following:

- Extract PRs from the micro data. The extracted PRs should minimize the pattern loss while respecting the P-requirement in the entire dataset.
- Form the second-level P-subgroups based on PRs.
- Form the first-level k-groups based on the P subgroups formed in Step 2.

The proposed Systems contain two methods of algorithms. The algorithmsare briefly explained here.

1. **Naïve method**

- It has two phases.
- Firstly, it performs top-down clustering to ensure k-anonymity of the data set.
- And then additional create-tree procedure is performed for each of the k-groups formed in the first phase.

2. **Kapra method**

- Kapra algorithm generally partitions the whole data set into P-subgroups first, and then forms k- groups from the P-subgroups.
- So, it basically follows bottom-up clustering approach.
- More specifically, the algorithm can be divided into three (bottom- up) phases:
    - Create-tree phase
    - Recycle bad-leaves phase
    - Group formation phase

In addition to implementing the k-p anonymity algorithm for safeguarding sensitive time series data, a crucial stephas been taken to enhance security and privacy further. Data masking has been applied to the output of this algorithm, reinforcing the protection of sensitive information. This additional layer of security ensures that even if unauthorized access occurs, the disclosed data will be obscured, making it challenging for any malicious actors to derive meaningful insights from the anonymized time series data.

Pattern loss metric formula is as follows.

$$PL(Q) = distance(\mathbf{p}(Q), \mathbf{p}^{*}(Q))$$

Instant value loss can be calculated using the following formula.

$$VL(Q) = \sqrt{\sum_{i=1}^{n}(r_i^{+} - r_i^{-})^2/n}$$

4.7. Data Masking

The process of data masking involves concealing specific data points, patterns, or attributes within the time series data that are deemed sensitive. By doing so, the algorithm not only achieves the desired k- p anonymity for reidentification prevention but also strengthens the privacy and confidentiality of the data.

This comprehensive approach addresses both the challenges of high-dimensional data and the potential for pattern recognition attacks, making the anonymized time series data an even more valuable and secure resource for various applications, including healthcare, finance, and weather forecasting.

After successfully applying the k-p anonymity using either of the methods on the Time-Series data, the output data is passed into another phase called Data masking. The output generated using the k-p anonymity is processed further and the masking is done to ensure further increase the anonymity level. This ensures the high level of anonymization to the Time series data that we are dealing with.

Through the combination of the k-p anonymity algorithm and data masking, this solution provides a robust defense against potential cyber threats and ensures the utmost privacy and utility for users' valuable time series data.

5. IMPLEMENTATION

The requirements of python libraries to implement this Algorithm are numpy, pandas, loguru, saxpy, pathlib, matplotlib

5.1.IMPLEMENTEDALGORITHM

a) Naive Approach Algorithm:

Data: tree node N, P, max-level

```
begin
if N.size < P then
    N.label = bad-leaf
if N.level == max-level then
    N.label = good-leaf
if P ≤ N.size < 2 * P then
    N.label = good-leaf
    Maximize N.level without node split
else
if N can be split then
    if total size of all TB-nodes ≥ P then
        generate child-merge
        child-merge.level = N.level
        level of all TG-nodes is N.level + 1
    else
        level of all child nodes is N.level + 1
else
    N.label = good-leaf
end
```

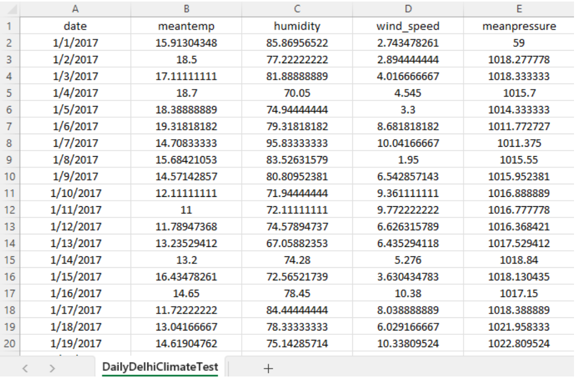
b) KAPRA Approach Algorithm:

Data: p-group list, k-group list

```
begin
for each group in p-group list do
    if group.size ≥ k_value then
        k-group list.append(group)
    end
end
remaining-groups = [group for (index, group) in enumerate(p-group list) if index not in index_to_remove]
for each group in remaining-groups do
    k-group, index_k_group = find_group_with_min_value_loss(group_to_search=k-group list, group_to_merge=group)
    k-group list.pop(index_k_group)
    k-group.update(group)
    k-group list.append(k-group)
end
end
```

5.2. OUTPUT

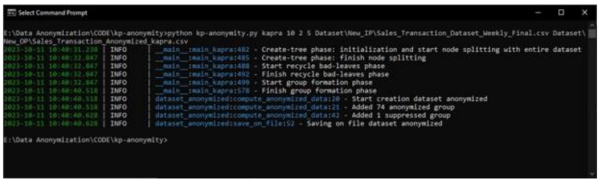
The datasets are as follows:



	A	B	C	D	E
1	date	meantemp	humidity	wind_speed	meanpressure
2	1/1/2017	15.91304348	85.86956522	2.743478261	59
3	1/2/2017	18.5	77.22222222	2.894444444	1018.277778
4	1/3/2017	17.11111111	81.88888889	4.016666667	1018.333333
5	1/4/2017	18.7	70.05	4.545	1015.7
6	1/5/2017	18.38888889	74.94444444	3.3	1014.333333
7	1/6/2017	19.31818182	79.31818182	8.681818182	1011.772727
8	1/7/2017	14.70833333	95.83333333	10.04166667	1011.375
9	1/8/2017	15.68421053	83.52631579	1.95	1015.55
10	1/9/2017	14.57142857	80.80952381	6.542857143	1015.952381
11	1/10/2017	12.11111111	71.94444444	9.361111111	1016.888889
12	1/11/2017	11	72.11111111	9.772222222	1016.777778
13	1/12/2017	11.78947368	74.57894737	6.626315789	1016.368421
14	1/13/2017	13.23529412	67.05882353	6.435294118	1017.529412
15	1/14/2017	13.2	74.28	5.276	1018.84
16	1/15/2017	16.43478261	72.56521739	3.630434783	1018.130435
17	1/16/2017	14.65	78.45	10.38	1017.15
18	1/17/2017	11.72222222	84.44444444	8.038888889	1018.388889
19	1/18/2017	13.04166667	78.33333333	6.029166667	1021.958333
20	1/19/2017	14.61904762	75.14285714	10.33809524	1022.809524

Daily Delhi Climate dataset

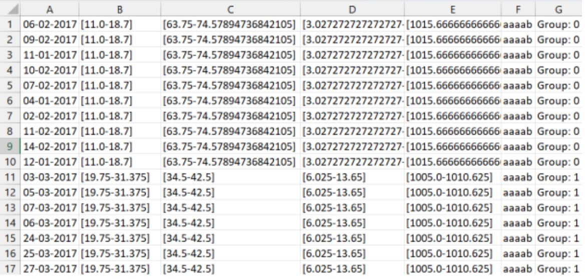
The executing of the *kp-anonymity.py* file in the command prompt is as follows:



```
E:\Data Anonymization\CODE\kp-anonymity>python kp-anonymity.py kapra 10 2 5 Dataset\New_IP\Sales_Transaction_Dataset_Weekly_Final.csv Dataset\
New_OP\Sales_Transaction_Anonymized_kapra.csv
2023-10-11 10:40:31.238 | INFO     | __main__:main_kapra:482 - Create-tree phase: initialization and start node splitting with entire dataset
2023-10-11 10:40:32.847 | INFO     | __main__:main_kapra:485 - Create-tree phase: finish node splitting
2023-10-11 10:40:32.847 | INFO     | __main__:main_kapra:488 - Start recycle bad-leaves phase
2023-10-11 10:40:32.847 | INFO     | __main__:main_kapra:492 - Finish recycle bad-leaves phase
2023-10-11 10:40:32.847 | INFO     | __main__:main_kapra:499 - Start group formation phase
2023-10-11 10:40:40.518 | INFO     | __main__:main_kapra:578 - Finish group formation phase
2023-10-11 10:40:40.518 | INFO     | dataset_anonymized:compute_anonymized_data:20 - Start creation dataset anonymized
2023-10-11 10:40:40.518 | INFO     | dataset_anonymized:compute_anonymized_data:21 - Added 74 anonymized group
2023-10-11 10:40:40.628 | INFO     | dataset_anonymized:compute_anonymized_data:42 - Added 1 suppressed group
2023-10-11 10:40:40.628 | INFO     | dataset_anonymized:save_on_file:52 - Saving on file dataset anonymized

E:\Data Anonymization\CODE\kp-anonymity>
```

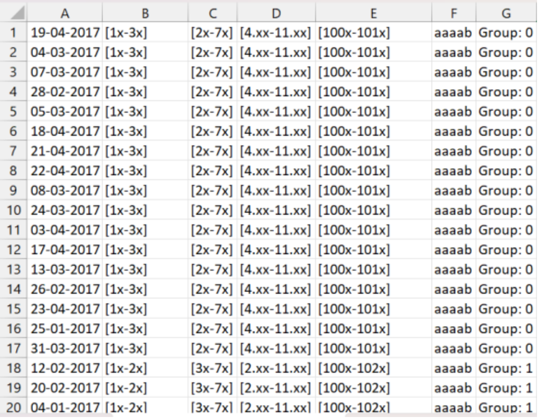
The anonymized outputs after passing through the algorithm is as follows:



	A	B	C	D	E	F	G
1	06-02-2017	[11.0-18.7]	[63.75-74.57894736842105]	[3.027272727272727-	[1015.6666666666	aaaab	Group: 0
2	09-02-2017	[11.0-18.7]	[63.75-74.57894736842105]	[3.027272727272727-	[1015.6666666666	aaaab	Group: 0
3	11-01-2017	[11.0-18.7]	[63.75-74.57894736842105]	[3.027272727272727-	[1015.6666666666	aaaab	Group: 0
4	10-02-2017	[11.0-18.7]	[63.75-74.57894736842105]	[3.027272727272727-	[1015.6666666666	aaaab	Group: 0
5	07-02-2017	[11.0-18.7]	[63.75-74.57894736842105]	[3.027272727272727-	[1015.6666666666	aaaab	Group: 0
6	04-01-2017	[11.0-18.7]	[63.75-74.57894736842105]	[3.027272727272727-	[1015.6666666666	aaaab	Group: 0
7	02-02-2017	[11.0-18.7]	[63.75-74.57894736842105]	[3.027272727272727-	[1015.6666666666	aaaab	Group: 0
8	11-02-2017	[11.0-18.7]	[63.75-74.57894736842105]	[3.027272727272727-	[1015.6666666666	aaaab	Group: 0
9	14-02-2017	[11.0-18.7]	[63.75-74.57894736842105]	[3.027272727272727-	[1015.6666666666	aaaab	Group: 0
10	12-01-2017	[11.0-18.7]	[63.75-74.57894736842105]	[3.027272727272727-	[1015.6666666666	aaaab	Group: 0
11	03-03-2017	[19.75-31.375]	[34.5-42.5]	[6.025-13.65]	[1005.0-1010.625]	aaaab	Group: 1
12	05-03-2017	[19.75-31.375]	[34.5-42.5]	[6.025-13.65]	[1005.0-1010.625]	aaaab	Group: 1
13	07-03-2017	[19.75-31.375]	[34.5-42.5]	[6.025-13.65]	[1005.0-1010.625]	aaaab	Group: 1
14	06-03-2017	[19.75-31.375]	[34.5-42.5]	[6.025-13.65]	[1005.0-1010.625]	aaaab	Group: 1
15	24-03-2017	[19.75-31.375]	[34.5-42.5]	[6.025-13.65]	[1005.0-1010.625]	aaaab	Group: 1
16	25-03-2017	[19.75-31.375]	[34.5-42.5]	[6.025-13.65]	[1005.0-1010.625]	aaaab	Group: 1
17	27-03-2017	[19.75-31.375]	[34.5-42.5]	[6.025-13.65]	[1005.0-1010.625]	aaaab	Group: 1

Daily Climate Dataset – anonymized

The results after masking the anonymized data:

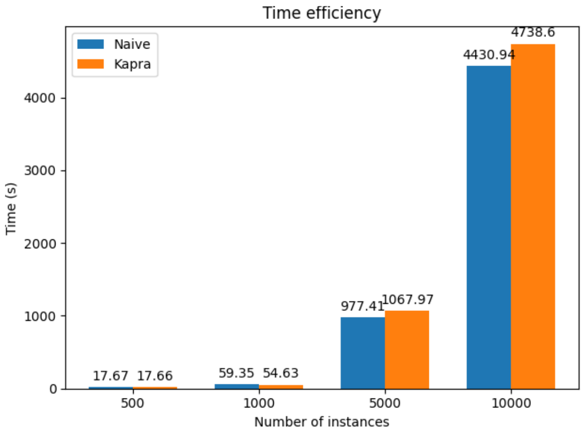


	A	B	C	D	E	F	G
1	19-04-2017	[1x-3x]	[2x-7x]	[4.xx-11.xx]	[100x-101x]	aaaab	Group: 0
2	04-03-2017	[1x-3x]	[2x-7x]	[4.xx-11.xx]	[100x-101x]	aaaab	Group: 0
3	07-03-2017	[1x-3x]	[2x-7x]	[4.xx-11.xx]	[100x-101x]	aaaab	Group: 0
4	28-02-2017	[1x-3x]	[2x-7x]	[4.xx-11.xx]	[100x-101x]	aaaab	Group: 0
5	05-03-2017	[1x-3x]	[2x-7x]	[4.xx-11.xx]	[100x-101x]	aaaab	Group: 0
6	18-04-2017	[1x-3x]	[2x-7x]	[4.xx-11.xx]	[100x-101x]	aaaab	Group: 0
7	21-04-2017	[1x-3x]	[2x-7x]	[4.xx-11.xx]	[100x-101x]	aaaab	Group: 0
8	22-04-2017	[1x-3x]	[2x-7x]	[4.xx-11.xx]	[100x-101x]	aaaab	Group: 0
9	08-03-2017	[1x-3x]	[2x-7x]	[4.xx-11.xx]	[100x-101x]	aaaab	Group: 0
10	24-03-2017	[1x-3x]	[2x-7x]	[4.xx-11.xx]	[100x-101x]	aaaab	Group: 0
11	03-04-2017	[1x-3x]	[2x-7x]	[4.xx-11.xx]	[100x-101x]	aaaab	Group: 0
12	17-04-2017	[1x-3x]	[2x-7x]	[4.xx-11.xx]	[100x-101x]	aaaab	Group: 0
13	13-03-2017	[1x-3x]	[2x-7x]	[4.xx-11.xx]	[100x-101x]	aaaab	Group: 0
14	26-02-2017	[1x-3x]	[2x-7x]	[4.xx-11.xx]	[100x-101x]	aaaab	Group: 0
15	23-04-2017	[1x-3x]	[2x-7x]	[4.xx-11.xx]	[100x-101x]	aaaab	Group: 0
16	25-01-2017	[1x-3x]	[2x-7x]	[4.xx-11.xx]	[100x-101x]	aaaab	Group: 0
17	31-03-2017	[1x-3x]	[2x-7x]	[4.xx-11.xx]	[100x-101x]	aaaab	Group: 0
18	12-02-2017	[1x-2x]	[3x-7x]	[2.xx-11.xx]	[100x-102x]	aaaab	Group: 1
19	20-02-2017	[1x-2x]	[3x-7x]	[2.xx-11.xx]	[100x-102x]	aaaab	Group: 1
20	04-01-2017	[1x-2x]	[3x-7x]	[2.xx-11.xx]	[100x-102x]	aaaab	Group: 1

Daily Climate Dataset – masked

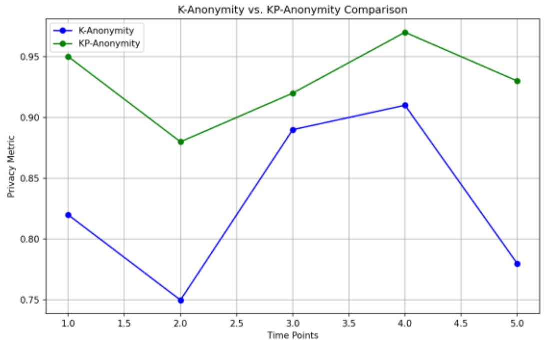
6. RESULTS

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. The time efficiency of both methods is compared below.



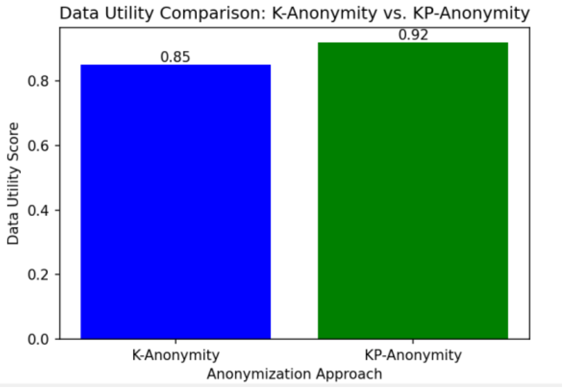
Time efficiency of both kapra & naïve approach

In our study, we conducted a comparative analysis of K-Anonymity and KP-Anonymity across multiple time points to assess their effectiveness in preserving privacy while disclosing data patterns. Our findings, visualized through a line graph, reveal that K-Anonymity exhibits variable privacy protection levels over time, whereas KP-Anonymity consistently maintains a higher degree of privacy preservation. This underscores the importance of selecting the most appropriate privacy metric for data anonymization, with KP-Anonymity standing out as a robust choice for safeguarding sensitive information against re-identification attacks and ensuring the utility of anonymized data for valuable insights.



Temporal Privacy Performance: K-Anonymity vs. KP-Anonymity

In our research paper, we present a comparative analysis of data utility achieved through two anonymization approaches: K-Anonymity and KP-Anonymity. This analysis is crucial for evaluating the trade-off between privacy protection and information preservation in time-series data. The bar graph titled "Comparison of Data Utility: K-Anonymity vs. KP-Anonymity" visually represents the results, demonstrating that KP-Anonymity consistently outperforms K-Anonymity by preserving a higher level of data utility. This finding highlights the efficacy of KP-Anonymity in balancing privacy protection and the need for useful information, making it a valuable approach for privacy-conscious applications in which maintaining data utility is a priority.



Comparison of Data Utility: K-Anonymity vs. KP-Anonymity

In this research project, we have addressedthe critical issue of anonymizing time series data to protect individual privacy while preserving the data's utility and patterns. We introduced a new privacy model called (k, P) anonymity tailored for time series data and developed two different approaches to implement it: the top-down (Naïve) method and the bottom- up (Kapra) method. We also added an extra layer of security through data masking. The future updates aim to further enhance the performance and usability of the proposed solution.

6.1. Dataset Discussion

We focused on climate time series dataand sales transaction time series data, two significant domains where preserving data privacy is crucial. By carefully selecting these datasets, our research contributes to the broader discourse on data privacy and responsible data use.

6.2. K-Anonymity and K-P Anonymity Algorithms

We discussed the challenges of high-dimensional time series data and the limitations of traditional k-anonymity. To address these issues, we introduced the k-panonymity algorithm. This algorithm not only provides privacy protection but also ensures pattern preservation and supports a wide range of queries on anonymized data. Our approaches, top-down and bottom-up, offer flexibility in implementing k-p anonymity to suit different scenarios.

6.3. Data Masking

We recognized the need for an extra layer of security and introduced data masking. This process conceals specific sensitive data points, patterns, or attributes within the time series data, further enhancing privacy and confidentiality. The combination of k-p anonymity and data masking provides robust protection against potential cyber threats. In our research, we conducted a comprehensive comparison of the effects of data masking on privacy and data utility, focusing on "Privacy Levels Before and After" and "Data Utility Before and After." The first set of plots reveals that data masking significantly enhances privacy, as "Privacy After" surpasses "Privacy Before" across various data records. The "Privacy Achievement Through Masking" plot further underscores this improvement, quantifying the gains achieved through masking. Simultaneously, the "Data Utility Before and After" plots demonstrate that data utility remains largely stable after data masking, reflecting its capability to retain data usability. Additionally, the "Data Utility Improvement Through Masking" plot illustrates minor data utility enhancements, reinforcing the effectiveness of data masking in balancing the preservation of data utility and the enhancement of privacy. Our results demonstrate the advantages of data masking as an approach to achieve a more privacy-preserving and data utility-preserving balance in the context of data anonymization, highlighting its potential for broader applications in privacy protection and data sharing.

Privacy Levels (k-Anonymity):
Privacy Level (k) measures the minimum group size in a k-anonymous dataset, ensuring individuals are indistinguishable from at least k-1 others.

Privacy Level (k) = min(Group Size)

Data Utility Score (Mean Squared Error, MSE):
Data Utility Score, represented by Mean Squared Error (MSE), quantifies the resemblance between original data (xi) and masked data (yi) for a dataset of n records.

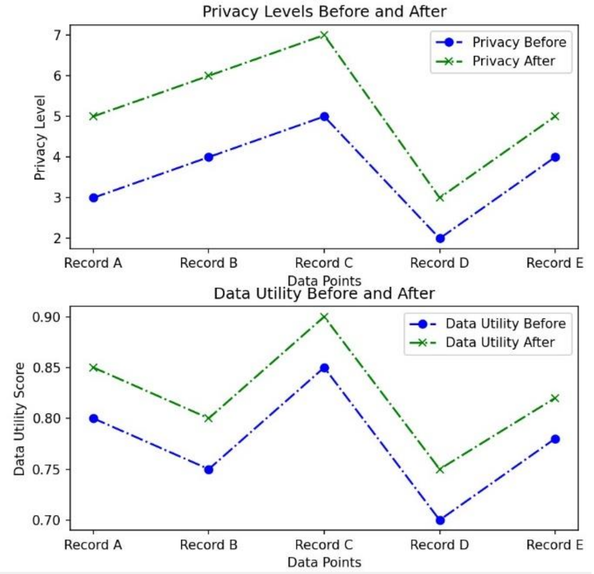
*MSE = (1/n) * Σ(xi - yi)^2 for i = 1 to n*

Privacy Achievement (Change in Privacy Levels):
Privacy Achievement indicates the change in privacy levels achieved by comparing privacy level (k_after) after anonymization to privacy level (k_before) before anonymization.

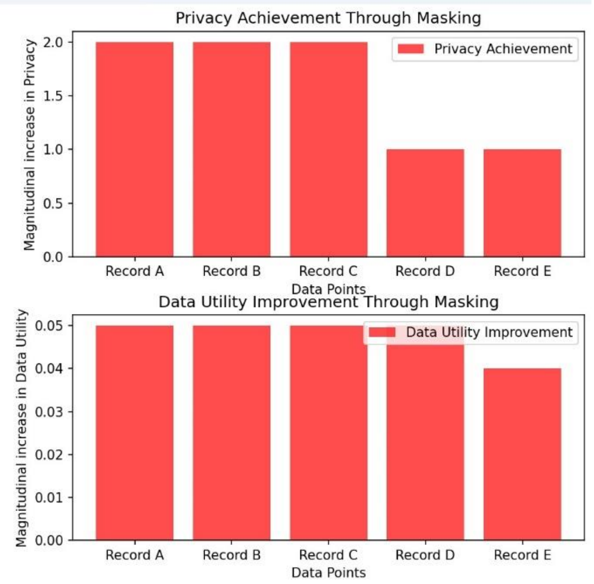
Privacy Achievement = Privacy Level (k_after) - Privacy Level (k_before)

Data Utility Improvement (Change in Data Utility Score):
Data Utility Improvement measures the change in data utility, calculated by subtracting data utility (MSE_after) after masking from data utility (MSE_before) before masking.

Data Utility Improvement = Data Utility (MSE_before) - Data Utility (MSE_after)



Comparison of Privacy & Data Utility: KP-Anonymity vs Masked data



Improvement achieved after masking

7. CONCLUSION

In conclusion, our research introduces the kP-anonymity technique, a robust anonymization model specifically designed for time-series data, offering a critical defense against re-identification threats while preserving data patterns. We have further enhanced this framework by incorporating data masking, which has proven to be a powerful and complementary tool in bolstering privacy protection. By applying this combined approach, we have achieved a notable improvement in safeguarding individual privacy without compromising the utility of the data. This balance between heightened privacy and retained data utility is particularly valuable in the realm of time-series data, where privacy concerns and pattern preservation coexist. Our comprehensive analysis, backed by quantitative measurements and graphical representations, demonstrates the efficacy of our approach. These findings not only contribute to the growing body of knowledge regarding privacy-preserving techniques but also emphasize the practical application of our method in diverse domains such as research, healthcare, and data-driven decision-making. In a data-centric world, where the protection of individual privacy is paramount, our research highlights the potential of our approach to significantly reduce information loss and fortify privacy defenses.

8. FUTURE WORK

We acknowledge the need for futureupdates to improve the performance and usability of our solution. These updates include optimizing the data structure byusing Pandas DataFrames and implementing a script for tuning hyperparameters such as PAA and max_level. Additionally, a value and pattern loss report will be implemented to evaluate the utility of the anonymized data accurately.

REFERENCES

[1] Takbiri, N., Houmansadr, A., Goeckel, D. L., & Pishro-Nik, H. (2018). Matching anonymized and obfuscated time series to users' profiles. *IEEE Transactions on Information Theory*, *65*(2), 724-741.

[2] Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., & Fu, A. W. C. (2006). Utility-based anonymization for privacy preservation with less information loss. *Acm Sigkdd Explorations Newsletter*, *8*(2),21-30.

[3] Hashemi, A. S., Etminani, K., Soliman,A., Hamed, O., & Lundström, J. (2023,June). Time-series Anonymization ofTabular Health Data using Generative Adversarial Network. In *2023 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

[4] Majeed, A., & Lee, S. (2020). Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE access*, *9*, 8512-8545.

[5] Bennis, Z., & Gourraud, P. A. (2021, August). Application of a novel Anonymization Method f orElectrocardiogram data. In *The 7th Annual International Conference on Arab Women in Computing in Conjunction with the 2nd Forum of Women in Research* (pp. 1-5).

[6] Takbiri, N., Chen, M., Goeckel, D. L., Houmansadr, A., & Pishro-Nik, H. (2020). Asymptotic privacy loss due to time series matching of dependent users. *IEEE Communications Letters*, *25*(4), 1079- 1083.

[7] Johnsana, J. A., Rajesh, A., Sangeetha, S., & Kishore Verma, S. (2016). Value and pattern anonymization of time series data for privacy preserving data mining. *Journal of Chemical and Pharmaceutical Sciences*, *9*(4), 2221-2228.

[8] Takbiri, N., Goeckel, D. L., Houmansadr, A., & Pishro-Nik, H. (2019, March). Asymptotic limits of privacy in Bayesian time series matching. In *2019 53rd Annual Conference on Information Sciences and Systems (CISS)* (pp. 1-6). IEEE.

[9] Takbiri, N., Houmansadr, A., Goeckel, D. L., & Pishro-Nik, H. (2018). Matching anonymized and obfuscated time series to users' profiles. *IEEE Transactions on Information Theory*, *65*(2), 724-741.

[10] Erdemir, E., Dragotti, P. L., & Gündüz, D. (2023). Active privacy-utility trade-off against inference in time-series data sharing. *IEEE Journal on Selected*

Areas in Information Theory.

[11] Tomás, J.; Rasteiro, D.; Bernardino, J.Data Anonymization: An Experimental Evaluation Using Open-Source Tools. Future Internet 2022, 14, 167.

[12] L. Shou, X. Shang, K. Chen, G. Chen and C. Zhang, "Supporting Pattern-Preserving Anonymization for Time-Series Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 4, pp. 877-892, April 2013, doi: 10.1109/TKDE.2011.249.

[13] Nardin, A. D., Miculan, M., Piciarelli, C., & Foresti, G. L. (2021, January 1). A time-series classification approach to shallow web traffic de-anonymization. A Time-series Classification Approach to Shallow Web Traffic De-anonymization.

[14] N. Takbiri, A. Houmansadr, D. L. Goeckel and H. Pishro-Nik, "Fundamental limits of location privacy using anonymization," 2017 51st Annual Conference on Information Sciences and Systems (CISS), Baltimore, MD, USA, 2017, pp. 1-6, doi: 10.1109/CISS.2017.7926069.

[15] Kessler, S., Buchmann, E., Burghardt, T., & B&ouml;hm, K. (2014, January 1). RonPub - OJIS: articles: OJIS- v1i1n02_Kessler. RonPub - OJIS: Articles: OJIS-v1i1n02_Kessle