BCI2001

DATA PRIVACY

(FALL SEMESTER 2022 - 23)

(SLOT: B2+TB2)

Final Project Review Document

Anonymisation of User's Time Series Data

Under the guidance of

JASMIN T JOSE

Associate Professor Grade - 1

Submitted by

Aravinth M – 20BCI0192

Keerthivasan K – 20BCI0193

Poovarasan A- 20BCI0194

Pratheeshkumar N-20BCI0195

Mukundhan D-20BCI0291

B.Tech. in Computer Science and Engineering(SCOPE)

VIT

Vellore Institute of Technology

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

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.

Keywords: Time series data, Anonymisation algorithms, k-anonymisation, k-p anonymisation, high utility of anonymised data, high privacy of anonymised data, anonymisation of time series data.

1. Introduction

1.1. Theoretical Background

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.

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

Figure 1 - Example of Time Series Data

1.2. Motivation

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

1.3. Aim of the Proposed Work

The main aim of this work would be proposing a new algorithm for anonymising the 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 by maintaining the statistical properties of data.

1.4. Objective of the Proposed Work

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

2. Literature Survey

2.1. Survey of Existing Work

S. No	**Name**	**Year**	**Journal**	**Author(s)**	**Technique/ Algorithm Used**	**Limitations**
1	Supporting Pattern Preserving Anonymizatio n 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 kanonymity model cannot effectively address this problem as it may suffer severe pattern loss. Proposed a	The current solution imposes a very strict constraint on PR equality and this may cause serious pattern loss.

					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.	
2	Utility-Based Anonymization for Privacy Preservation with Less Information Loss	2021	ACM	Jian Xu1 Wei Wang1 Jian Pei2 Xiaoyuan Wang1 Baile Shi1 Ada WaiChee 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 utilitybased anonymization were developed. The bottom-up method and the	The computation time is often a secondary consideration yielding to the quality.

					top-down method achieve better anonymization than the MultiDim.	
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 tradeoff 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	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	Implementation of the anonymizing summarization involves shape preserving greedy elimination and

					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	aggregation that supports parallel cluster processing for big data implementation.
5	Pattern sensitive Time	2014	IEEE	Stephan Kessler,	Time series anonymization	They Proposed (n,l,k)

	series Anonymizatio and its Application to Energy Consumption Data			Erik Buchmann, Thorben Burghardt, Klemens Bo ̈hm	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	Anonymity To reduce the Information loss, But it does not work properly for univariate series
6	Supporting Pattern - Preserving Anonymizatio n 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	This model publishes both the attribute values and the patterns of time series in separate data forms.

					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.	
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	In this paper a combination of novel methodologies Kanonymization (SKY), Symbolic polynomial with

					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	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.
9	This Algorithm provides low entropy information	2019	IEEE	Erik Wik	In this thesis, an investigation is made into how to protect	The results show that PC - KAPRA offers a large

	loss				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	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,	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

					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.	proved both for the i.i.d. case as well as the more general Markov chain mode

2.2. Summary/Gaps Identified in the Survey

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

i. 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.

ii. 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.

iii. 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 everyone 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.

iv. 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

v. 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.

vi. 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 reidentification attack will happen. Few of the research papers where privacy achieved is very high, has lots of information loss.

vii. Maintaining variable privacy utility threshold depending upon the priority of healthcare 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 literatures 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 were multi-dimensional sparse data.

3. Overview of Proposed System

3.1. Introduction and Related Concepts

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 to high protection or low protection, thus resulting in poor utility (usage).

(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 a wide range of queries on anonymized data.

3.2. Framework, Architecture or Module for the Proposed System

(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 a wide range of queries on anonymized data. We have two algorithms for (k, P)-anonymity on time-series data. This anonymity model 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:

1. 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 model 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 model ensures anonymity on two levels. On the first level, the QI attributes are generalized to fulfill 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.

2. 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 kgroup. 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 global optimal solution requires combinatorial computation cost. Therefore, we will consider more efficient near-optimal solutions in the sequel.

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.

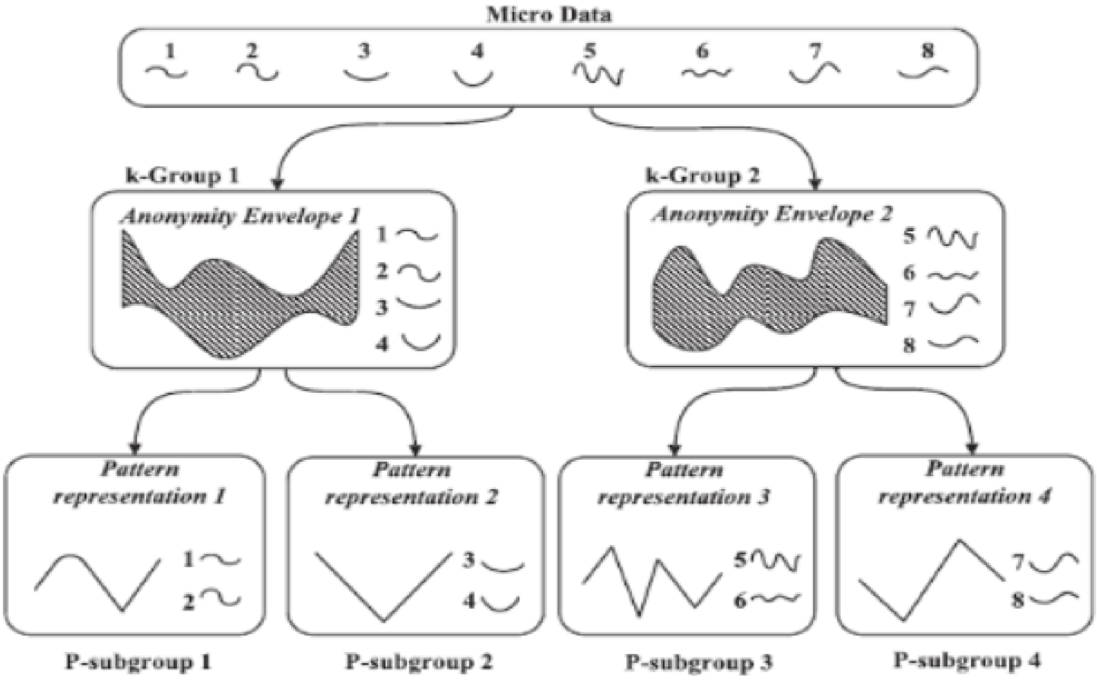
Step 2 is a challenging task and highly dependent on the PR form being used. Different PR forms may lead to very different implementations of this step. In whatever forms, the granularity of PR should be carefully tuned to achieve the optimization target of this step.

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.

The bottom-up approach is described in the following:

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

3.3. **Proposed System Model**



4. **Proposed System Analysis and Design**

4.1. **Introduction**

<u>Video link:</u>

The project demonstration is done and the hyperlink for video is provided here.

https://drive.google.com/file/d/1XpqxFqa0Bx808blcq1t1HAc0y6CNJwLv/view?usp=share_link

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

<u>Algorithms</u>

- **<u>Naïve method</u>**
  - 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.
- **<u>Kapra method</u>**

- 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

<u>**Input - Dataset**</u>

For the implementation, we have taken several datasets and performed the anonymisation. Here, two of them have been shown. The sample of the datasets used for the implementation are as follows:

1) Daily Climate data of Delhi

	A	B	C	D	E
1	date	meantemp	humidity	wind_speed	meanpressure
2	2017-01-01	15.9130434782609	85.8695652173913	2.74347826086957	59
3	2017-01-02	18.5	77.2222222222222	2.89444444444444	1018.27777777778
4	2017-01-03	17.1111111111111	81.8888888888889	4.01666666666667	1018.33333333333
5	2017-01-04	18.7	70.05	4.545	1015.7
6	2017-01-05	18.3888888888889	74.9444444444444	3.3	1014.33333333333
7	2017-01-06	19.3181818181818	79.3181818181818	8.68181818181818	1011.77272727273
8	2017-01-07	14.7083333333333	95.8333333333333	10.0416666666667	1011.375
9	2017-01-08	15.6842105263158	83.5263157894737	1.95	1015.55
10	2017-01-09	14.5714285714286	80.8095238095238	6.54285714285714	1015.95238095238
11	2017-01-10	12.1111111111111	71.9444444444444	9.36111111111111	1016.88888888889
12	2017-01-11	11	72.1111111111111	9.77222222222222	1016.77777777778
13	2017-01-12	11.7894736842105	74.5789473684211	6.62631578947368	1016.36842105263
14	2017-01-13	13.2352941176471	67.0588235294118	6.43529411764706	1017.52941176471
15	2017-01-14	13.2	74.28	5.276	1018.84
16	2017-01-15	16.4347826086957	72.5652173913043	3.6304347826087	1018.13043478261
17	2017-01-16	14.65	78.45	10.38	1017.15
18	2017-01-17	11.7222222222222	84.4444444444444	8.03888888888889	1018.38888888889
19	2017-01-18	13.0416666666667	78.3333333333333	6.02916666666666	1021.95833333333
20	2017-01-19	14.6190476190476	75.1428571428571	10.3380952380952	1022.80952380952
21	2017-01-20	15.2631578947368	66.4736842105263	11.2263157894737	1021.78947368421
22	2017-01-21	15.3913043478261	70.8695652173913	13.695652173913	1020.47826086957
23	2017-01-22	18.44	76.24	5.868	1021.04
24	2017-01-23	18.1176470588235	76	6.75294117647059	1019.82352941176
25	2017-01-24	18.3478260869565	68.1304347826087	3.39130434782609	1018.86956521739
26	2017-01-25	21	69.96	8.756	1018.4
27	2017-01-26	16.1785714285714	91.6428571428571	8.46785714285714	1017.78571428571
28	2017-01-27	16.5	77.0416666666667	14.3583333333333	1018.125
29	2017-01-28	14.8636363636364	82.7727272727273	9.69090909090909	1019.63636363636

DailyDelhiClimateTest

2) Daily Confirmed Covid cases of Kerala

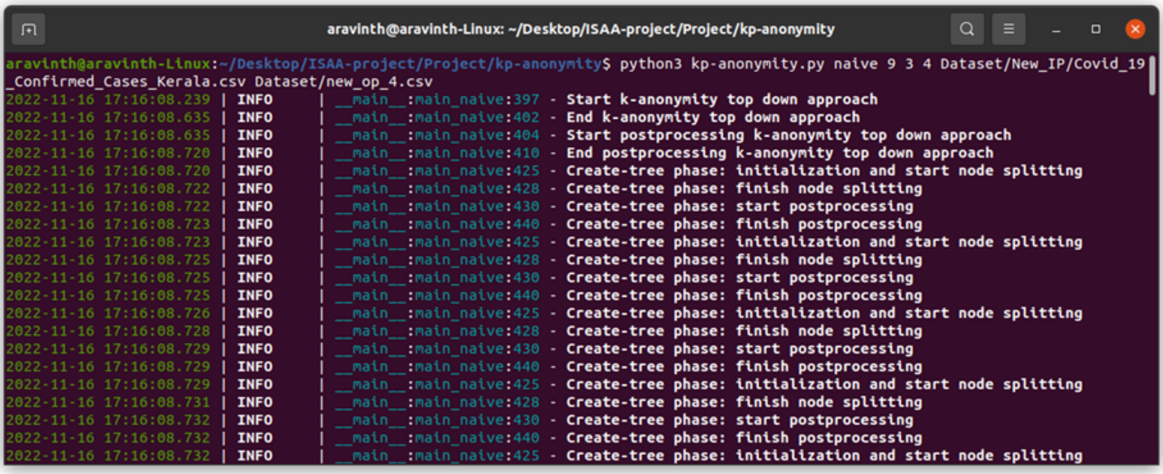
	A	B	C
1	Date	Confirmed	
2	2020-01-31	0	
3	2020-02-01	0	
4	2020-02-02	1	
5	2020-02-03	1	
6	2020-02-04	0	
7	2020-02-05	0	
8	2020-02-06	0	
9	2020-02-07	0	
10	2020-02-08	0	
11	2020-02-09	0	
12	2020-02-10	0	
13	2020-02-11	0	
14	2020-02-12	0	
15	2020-02-13	0	
16	2020-02-14	0	
17	2020-02-15	0	
18	2020-02-16	0	
19	2020-02-17	0	
20	2020-02-18	0	
21	2020-02-19	0	
22	2020-02-20	0	
23	2020-02-21	0	
24	2020-02-22	0	
25	2020-02-23	0	
26	2020-02-24	0	
27	2020-02-25	0	
28	2020-02-26	0	
29	2020-02-27	0	

Implementation - Anonymisation process

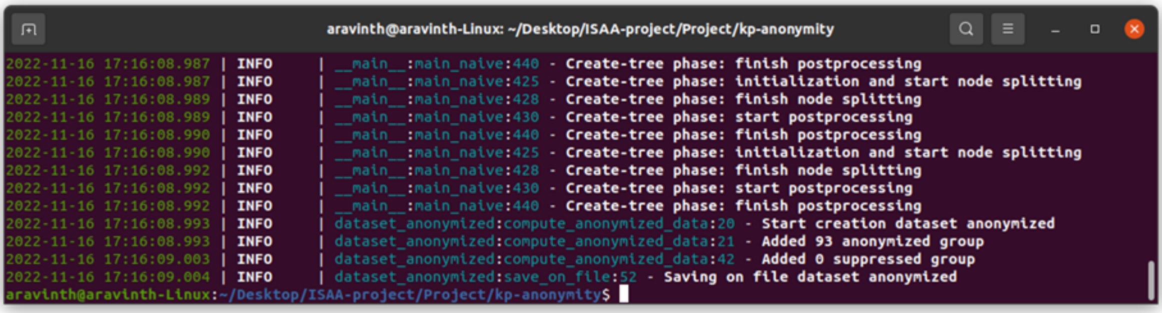
1) Anonymising - Daily Climate data of Delhi

```
aravinth@aravinth-Linux: ~/Desktop/ISAA-project/Project/kp-anonymity
aravinth@aravinth-Linux:~/Desktop/ISAA-project/Project/kp-anonymity$ python3 kp-anonymity.py kapra 10 2 5 Dataset/New_IP/
Covid_19_Confirmed_Cases_Kerala.csv Dataset/new_op_3.csv
2022-11-16 17:14:31.541 | INFO     | __main__:main_kapra:482 - Create-tree phase: initialization and start node splitting
 with entire dataset
2022-11-16 17:14:31.790 | INFO     | __main__:main_kapra:485 - Create-tree phase: finish node splitting
2022-11-16 17:14:31.790 | INFO     | __main__:main_kapra:488 - Start recycle bad-leaves phase
2022-11-16 17:14:31.790 | INFO     | __main__:main_kapra:492 - Finish recycle bad-leaves phase
2022-11-16 17:14:31.790 | INFO     | __main__:main_kapra:499 - Start group formation phase
2022-11-16 17:14:33.573 | INFO     | __main__:main_kapra:578 - Finish group formation phase
2022-11-16 17:14:33.573 | INFO     | dataset_anonymized:compute_anonymized_data:20 - Start creation dataset anonymized
2022-11-16 17:14:33.573 | INFO     | dataset_anonymized:compute_anonymized_data:21 - Added 67 anonymized group
2022-11-16 17:14:33.580 | INFO     | dataset_anonymized:compute_anonymized_data:42 - Added 0 suppressed group
2022-11-16 17:14:33.580 | INFO     | dataset_anonymized:save_on_file:52 - Saving on file dataset anonymized
aravinth@aravinth-Linux:~/Desktop/ISAA-project/Project/kp-anonymity$
```

2) Anonymising - Daily Confirmed Covid cases of Kerala



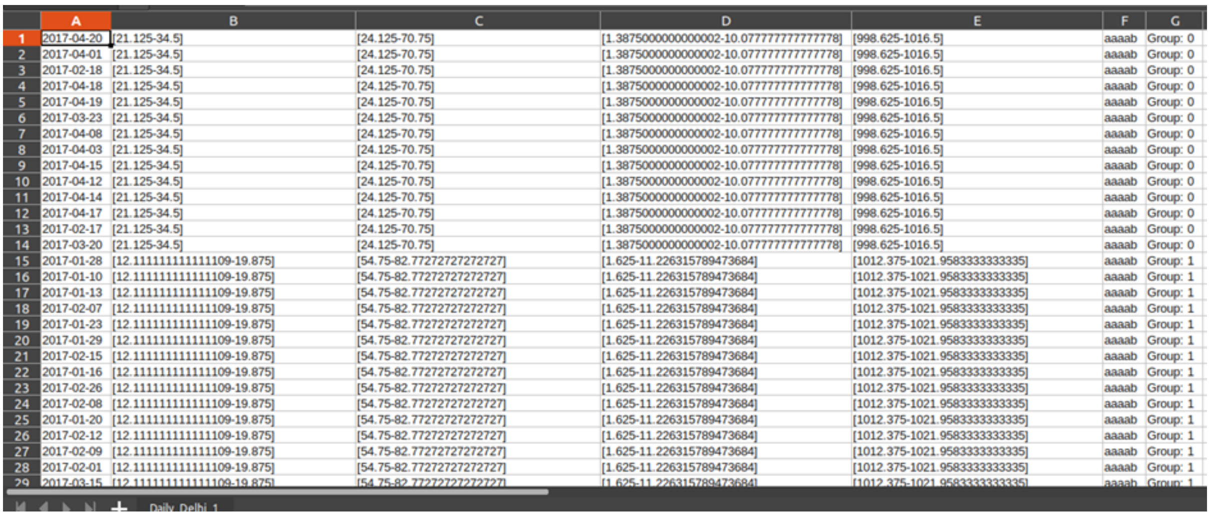
```
aravinth@aravinth-Linux: ~/Desktop/ISAA-project/Project/kp-anonymity
aravinth@aravinth-Linux:~/Desktop/ISAA-project/Project/kp-anonymity$ python3 kp-anonymity.py naive 9 3 4 Dataset/New_IP/Covid_19
_Confirmed_Cases_Kerala.csv Dataset/new_op_4.csv
2022-11-16 17:16:08.239 | INFO     | __main__:main_naive:397 - Start k-anonymity top down approach
2022-11-16 17:16:08.635 | INFO     | __main__:main_naive:402 - End k-anonymity top down approach
2022-11-16 17:16:08.635 | INFO     | __main__:main_naive:404 - Start postprocessing k-anonymity top down approach
2022-11-16 17:16:08.720 | INFO     | __main__:main_naive:410 - End postprocessing k-anonymity top down approach
2022-11-16 17:16:08.720 | INFO     | __main__:main_naive:425 - Create-tree phase: initialization and start node splitting
2022-11-16 17:16:08.722 | INFO     | __main__:main_naive:428 - Create-tree phase: finish node splitting
2022-11-16 17:16:08.722 | INFO     | __main__:main_naive:430 - Create-tree phase: start postprocessing
2022-11-16 17:16:08.723 | INFO     | __main__:main_naive:440 - Create-tree phase: finish postprocessing
2022-11-16 17:16:08.723 | INFO     | __main__:main_naive:425 - Create-tree phase: initialization and start node splitting
2022-11-16 17:16:08.725 | INFO     | __main__:main_naive:428 - Create-tree phase: finish node splitting
2022-11-16 17:16:08.725 | INFO     | __main__:main_naive:430 - Create-tree phase: start postprocessing
2022-11-16 17:16:08.725 | INFO     | __main__:main_naive:440 - Create-tree phase: finish postprocessing
2022-11-16 17:16:08.726 | INFO     | __main__:main_naive:425 - Create-tree phase: initialization and start node splitting
2022-11-16 17:16:08.728 | INFO     | __main__:main_naive:428 - Create-tree phase: finish node splitting
2022-11-16 17:16:08.729 | INFO     | __main__:main_naive:430 - Create-tree phase: start postprocessing
2022-11-16 17:16:08.729 | INFO     | __main__:main_naive:440 - Create-tree phase: finish postprocessing
2022-11-16 17:16:08.729 | INFO     | __main__:main_naive:425 - Create-tree phase: initialization and start node splitting
2022-11-16 17:16:08.731 | INFO     | __main__:main_naive:428 - Create-tree phase: finish node splitting
2022-11-16 17:16:08.732 | INFO     | __main__:main_naive:430 - Create-tree phase: start postprocessing
2022-11-16 17:16:08.732 | INFO     | __main__:main_naive:440 - Create-tree phase: finish postprocessing
2022-11-16 17:16:08.732 | INFO     | __main__:main_naive:425 - Create-tree phase: initialization and start node splitting
```



```
aravinth@aravinth-Linux: ~/Desktop/ISAA-project/Project/kp-anonymity
2022-11-16 17:16:08.987 | INFO     | __main__:main_naive:440 - Create-tree phase: finish postprocessing
2022-11-16 17:16:08.987 | INFO     | __main__:main_naive:425 - Create-tree phase: initialization and start node splitting
2022-11-16 17:16:08.989 | INFO     | __main__:main_naive:428 - Create-tree phase: finish node splitting
2022-11-16 17:16:08.989 | INFO     | __main__:main_naive:430 - Create-tree phase: start postprocessing
2022-11-16 17:16:08.990 | INFO     | __main__:main_naive:440 - Create-tree phase: finish postprocessing
2022-11-16 17:16:08.990 | INFO     | __main__:main_naive:425 - Create-tree phase: initialization and start node splitting
2022-11-16 17:16:08.992 | INFO     | __main__:main_naive:428 - Create-tree phase: finish node splitting
2022-11-16 17:16:08.992 | INFO     | __main__:main_naive:430 - Create-tree phase: start postprocessing
2022-11-16 17:16:08.992 | INFO     | __main__:main_naive:440 - Create-tree phase: finish postprocessing
2022-11-16 17:16:08.993 | INFO     | dataset_anonymized:compute_anonymized_data:20 - Start creation dataset anonymized
2022-11-16 17:16:08.993 | INFO     | dataset_anonymized:compute_anonymized_data:21 - Added 93 anonymized group
2022-11-16 17:16:09.003 | INFO     | dataset_anonymized:compute_anonymized_data:42 - Added 0 suppressed group
2022-11-16 17:16:09.004 | INFO     | dataset_anonymized:save_on_file:52 - Saving on file dataset anonymized
aravinth@aravinth-Linux:~/Desktop/ISAA-project/Project/kp-anonymity$
```

Output

1) Daily Climate data of Delhi - Anonymised



	A	B	C	D	E	F	G
1	2017-04-20	[21.125-34.5]	[24.125-70.75]	[1.3875000000000002-10.077777777777778]	[998.625-1016.5]	aaaab	Group: 0
2	2017-04-01	[21.125-34.5]	[24.125-70.75]	[1.3875000000000002-10.077777777777778]	[998.625-1016.5]	aaaab	Group: 0
3	2017-02-18	[21.125-34.5]	[24.125-70.75]	[1.3875000000000002-10.077777777777778]	[998.625-1016.5]	aaaab	Group: 0
4	2017-04-18	[21.125-34.5]	[24.125-70.75]	[1.3875000000000002-10.077777777777778]	[998.625-1016.5]	aaaab	Group: 0
5	2017-04-19	[21.125-34.5]	[24.125-70.75]	[1.3875000000000002-10.077777777777778]	[998.625-1016.5]	aaaab	Group: 0
6	2017-03-23	[21.125-34.5]	[24.125-70.75]	[1.3875000000000002-10.077777777777778]	[998.625-1016.5]	aaaab	Group: 0
7	2017-04-08	[21.125-34.5]	[24.125-70.75]	[1.3875000000000002-10.077777777777778]	[998.625-1016.5]	aaaab	Group: 0
8	2017-04-03	[21.125-34.5]	[24.125-70.75]	[1.3875000000000002-10.077777777777778]	[998.625-1016.5]	aaaab	Group: 0
9	2017-04-15	[21.125-34.5]	[24.125-70.75]	[1.3875000000000002-10.077777777777778]	[998.625-1016.5]	aaaab	Group: 0
10	2017-04-12	[21.125-34.5]	[24.125-70.75]	[1.3875000000000002-10.077777777777778]	[998.625-1016.5]	aaaab	Group: 0
11	2017-04-14	[21.125-34.5]	[24.125-70.75]	[1.3875000000000002-10.077777777777778]	[998.625-1016.5]	aaaab	Group: 0
12	2017-04-17	[21.125-34.5]	[24.125-70.75]	[1.3875000000000002-10.077777777777778]	[998.625-1016.5]	aaaab	Group: 0
13	2017-02-17	[21.125-34.5]	[24.125-70.75]	[1.3875000000000002-10.077777777777778]	[998.625-1016.5]	aaaab	Group: 0
14	2017-03-20	[21.125-34.5]	[24.125-70.75]	[1.3875000000000002-10.077777777777778]	[998.625-1016.5]	aaaab	Group: 0
15	2017-01-28	[12.111111111111109-19.875]	[54.75-82.77272727272727]	[1.625-11.226315789473684]	[1012.375-1021.9583333333335]	aaaab	Group: 1
16	2017-01-10	[12.111111111111109-19.875]	[54.75-82.77272727272727]	[1.625-11.226315789473684]	[1012.375-1021.9583333333335]	aaaab	Group: 1
17	2017-01-13	[12.111111111111109-19.875]	[54.75-82.77272727272727]	[1.625-11.226315789473684]	[1012.375-1021.9583333333335]	aaaab	Group: 1
18	2017-02-07	[12.111111111111109-19.875]	[54.75-82.77272727272727]	[1.625-11.226315789473684]	[1012.375-1021.9583333333335]	aaaab	Group: 1
19	2017-01-23	[12.111111111111109-19.875]	[54.75-82.77272727272727]	[1.625-11.226315789473684]	[1012.375-1021.9583333333335]	aaaab	Group: 1
20	2017-01-29	[12.111111111111109-19.875]	[54.75-82.77272727272727]	[1.625-11.226315789473684]	[1012.375-1021.9583333333335]	aaaab	Group: 1
21	2017-02-15	[12.111111111111109-19.875]	[54.75-82.77272727272727]	[1.625-11.226315789473684]	[1012.375-1021.9583333333335]	aaaab	Group: 1
22	2017-01-16	[12.111111111111109-19.875]	[54.75-82.77272727272727]	[1.625-11.226315789473684]	[1012.375-1021.9583333333335]	aaaab	Group: 1
23	2017-02-26	[12.111111111111109-19.875]	[54.75-82.77272727272727]	[1.625-11.226315789473684]	[1012.375-1021.9583333333335]	aaaab	Group: 1
24	2017-02-08	[12.111111111111109-19.875]	[54.75-82.77272727272727]	[1.625-11.226315789473684]	[1012.375-1021.9583333333335]	aaaab	Group: 1
25	2017-01-20	[12.111111111111109-19.875]	[54.75-82.77272727272727]	[1.625-11.226315789473684]	[1012.375-1021.9583333333335]	aaaab	Group: 1
26	2017-02-12	[12.111111111111109-19.875]	[54.75-82.77272727272727]	[1.625-11.226315789473684]	[1012.375-1021.9583333333335]	aaaab	Group: 1
27	2017-02-09	[12.111111111111109-19.875]	[54.75-82.77272727272727]	[1.625-11.226315789473684]	[1012.375-1021.9583333333335]	aaaab	Group: 1
28	2017-02-01	[12.111111111111109-19.875]	[54.75-82.77272727272727]	[1.625-11.226315789473684]	[1012.375-1021.9583333333335]	aaaab	Group: 1
29	2017-03-15	[12.111111111111109-19.875]	[54.75-82.77272727272727]	[1.625-11.226315789473684]	[1012.375-1021.9583333333335]	aaaab	Group: 1

Daily_Delhi_1

2) Daily Confirmed Covid cases of Kerala - Anonymised

	A	B	C	D
1	2020-02-25	[0.0-2.0]	aaaaa	Group: 0
2	2020-03-04	[0.0-2.0]	aaaaa	Group: 0
3	2020-02-11	[0.0-2.0]	aaaaa	Group: 0
4	2020-02-14	[0.0-2.0]	aaaaa	Group: 0
5	2020-03-19	[0.0-2.0]	bbbbb	Group: 0
6	2020-02-03	[0.0-2.0]	bbbbb	Group: 0
7	2020-05-08	[0.0-2.0]	bbbbb	Group: 0
8	2020-04-15	[0.0-2.0]	bbbbb	Group: 0
9	2020-05-02	[0.0-2.0]	bbbbb	Group: 0
10	2020-04-19	[0.0-2.0]	bbbbb	Group: 0
11	2022-05-16	[290.0-324.0]	bbbbb	Group: 1
12	2022-05-08	[290.0-324.0]	bbbbb	Group: 1
13	2022-04-21	[290.0-324.0]	bbbbb	Group: 1
14	2022-04-30	[290.0-324.0]	bbbbb	Group: 1
15	2022-05-15	[290.0-324.0]	bbbbb	Group: 1
16	2022-04-03	[290.0-324.0]	bbbbb	Group: 1
17	2022-04-13	[290.0-324.0]	bbbbb	Group: 1
18	2020-07-08	[290.0-324.0]	bbbbb	Group: 1
19	2022-04-24	[290.0-324.0]	bbbbb	Group: 1
20	2022-05-02	[290.0-324.0]	bbbbb	Group: 1
21	2020-05-13	[8.0-11.0]	bbbbb	Group: 2
22	2020-04-29	[8.0-11.0]	bbbbb	Group: 2
23	2020-04-11	[8.0-11.0]	bbbbb	Group: 2
24	2020-04-23	[8.0-11.0]	bbbbb	Group: 2
25	2020-04-08	[8.0-11.0]	bbbbb	Group: 2
26	2020-03-25	[8.0-11.0]	bbbbb	Group: 2
27	2020-04-07	[8.0-11.0]	bbbbb	Group: 2
28	2020-03-10	[8.0-11.0]	bbbbb	Group: 2
29	2020-04-22	[8.0-11.0]	bbbbb	Group: 2

<u>Outcome</u>

- ✓ Each value is changed into category
- ✓ The records are shuffled
- ✓ Same patterns in P-1 records of K records
- ✓ Patterns are preserved
- ✓ Utility (Usage) of data is increased
- ✓ All data are anonymized, with necessary privacy
- ✓ Each data record’s pattern is also displayed

- ✓ Each record is grouped by generalization

4.2. Requirement Analysis

4.2.1. Functional Requirements

4.2.1.1. Product Perspective

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. The perspective of this product would be improving the existing algorithmic techniques by introducing new algorithms and methods.

4.2.1.2. Product Features

This product (algorithm) features the protection of privacy of the user's various time series data such as healthcare data, financial transactions data, weather data, business accounts data etc., This product helps the anonymized data to be very useful to various tasks such as analysis, survey, senses, etc., by improving the usability (utility) of the data.

4.2.1.3. User Characteristics

The user(s) of this algorithm would be various organizations which have and use user's time series data for various purposes such as analyzing, testing, providing services to the clients, calculating usage analytics, etc.,

4.2.1.4. Assumption & Dependencies

Time series data is taken by noting the values for a particular attribute on different intervals. The data is a value of a particular attribute, noted frequently. The raw time series data is not anonymised and prone to various cyber-attacks. The raw time series data is then anonymised. The anonymised data is then used for various purposes.

4.2.1.5. Domain Requirements

This proposed product (algorithm) is most widely used in the areas wherever the values are recorded periodically in different time intervals. Some domains include

i) medical industry where the patient's heartbeat rate, blood glucose level, blood pressure level are taken as time series data,

ii) weather forecasting domain where the wind speed, humidity, temperature and pressure are noted periodically as time series data, and

iii) share market and investments related domain where the values of the stocks are noted periodically as time series data.

4.2.1.6. User Requirements

The end users of this algorithm (or software) can be classified into two different types.

i) The users (or organizations) apparently which use this software

ii) The users who give or contribute to give their data for various purposes.

Each user(s) has different requirements and purposes. The users (organizations) who uses the software explicitly requires the software to be more effective, with high performance and with as minimum run time as possible. And they require the final output i.e., the user’s anonymised data to be more usable (high utility). The end user(s) who contribute to provide the data, requires their data to be more privacy protected. In other words, they need not disclose their identity in any circumstances.

4.2.2. Non-Functional Requirements

4.2.2.1. Product Requirements

4.2.2.1.1. Efficiency

The efficiency of the algorithm is better than the existing algorithm k-anonymity. It is also analysed and shown here. The k-P anonymity algorithm is working effieciенty and produces the higher utility & privacy filled anonymised data.

4.2.2.1.2. Reliability

The reliability of the k-P anonymity algorithm is very well reliable because of it’s higher utility and privacy protection.

4.2.2.1.3. Portability

The portability of the algorithm (software) is obviously high because this algorithm can be embedded in any software and applications which is a real advantage for this kind of software. So it can be used for medical industry, financial industry, weather forecasting industry etc.,

4.2.2.1.4. Usability

The proposed k-P anonymity algorithm is very much useful and will be usable in anonymising the time series data effectively. Since the algorithm produces the well balanced anonymised data, the privacy as well as the utility (usability) of the anoymised data is eqally balanced. It keeps this algorithm usable in many cases.

4.2.2.2. **Organizational Requirements**

4.2.2.2.1. **Implementation Requirements (in terms of deployment)**

To implement this algorithm in the anonymisation tool, instead of using the already existing k-anonymity algorithm, the organization has 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 called ‘P’ should be introduced. That P variable should preserve the pattern of the time series data according to the privacy and utility requirements.

4.2.3. **System Requirements**

4.2.3.1. **Hardware Requirements**

The hardware requirements are classified into three categories.

i) Small dataset hardware requirements

Small datasets can easily be run in the basic personal computer having a minimum 4GB of RAM, 250GB of disk space, and basic CPU of intel i3 or equivalent AMD processor.

ii) Medium dataset hardware requirements

For running the medium datasets, a minimum of 8GB of RAM, 500GB of disk space, and a bit higher CPU of intel i5 or equivalent AMD processor.

iii) Large dataset hardware requirements

For running the large datasets of large user’s data in organizations, a minimum of 16GB of RAM, 1TB of disk space for the database, and a well-advanced CPU of intel i7 or equivalent AMD processor.

4.2.3.2. **Software Requirements**

The software requirements are mentioned here.

Operating system – Ubuntu Linux/ Debian Linux/ any distributions of Linux operating systems.

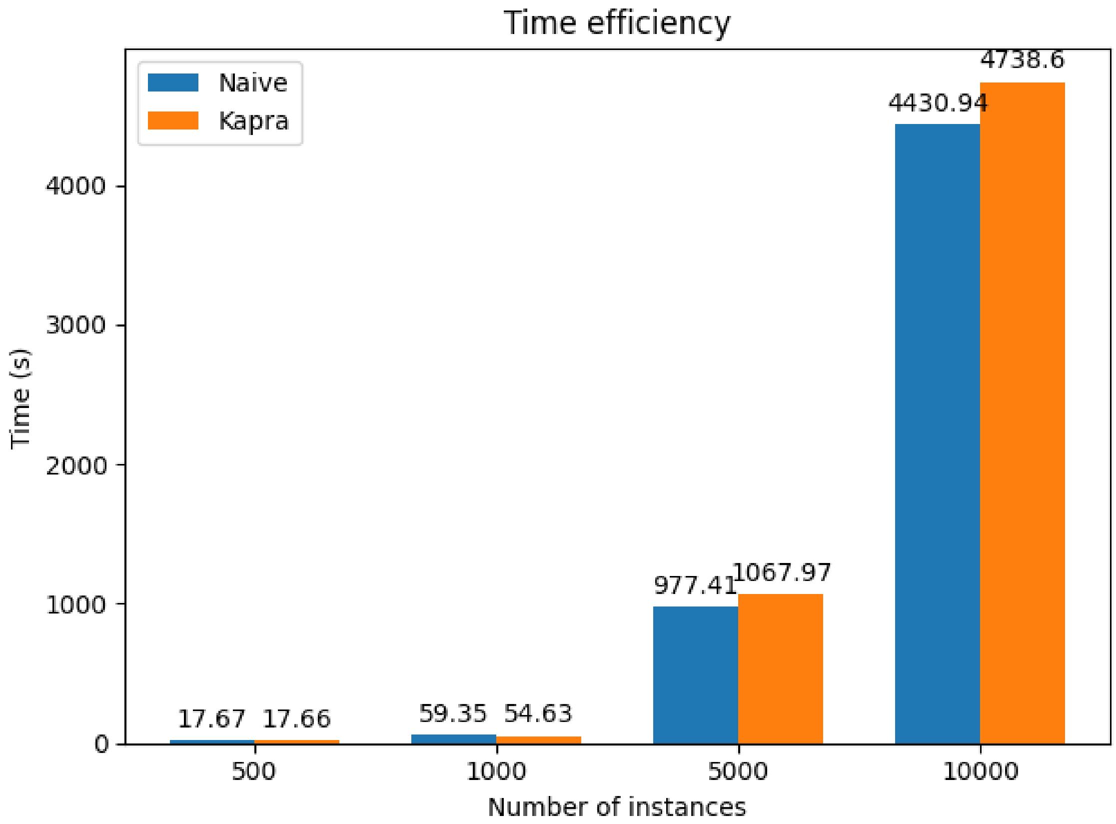
SQL Server – for accessing user’s time series from database

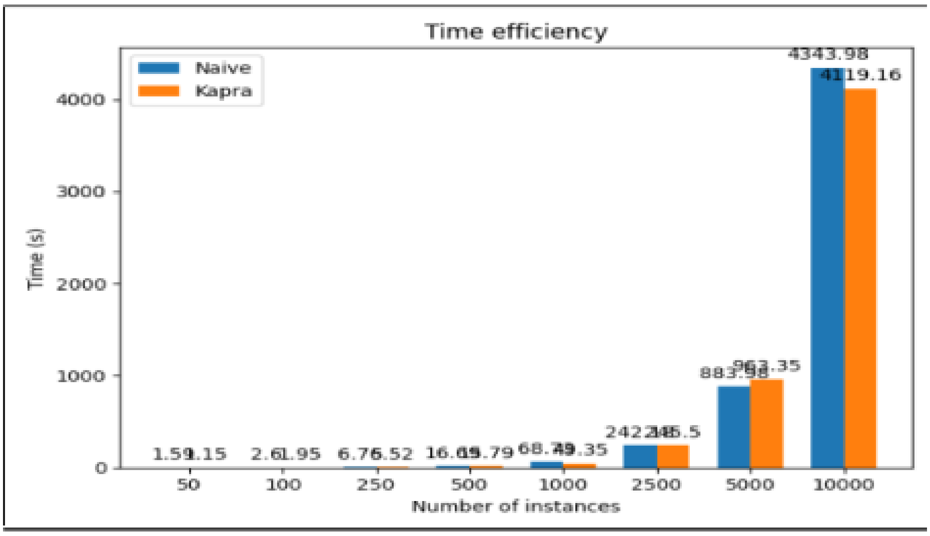
Python – any versions of python 3, most preferably the latest one

5. **Results and Discussion**

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.





6. 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] 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] Thouvenot, V., Nogues, D., & Gouttas, C. (2017). Data-driven Anonymization Process Applied to Time Series. In *SIMBig* (pp. 80-90).

[4] Ruta, D., Cen, L., & Damiani, E. (2015, October). Fast summarization and anonymization of multivariate big time series. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 1901-1904). IEEE.

[5] Kessler, S., Buchmann, E., Burghardt, T., & Böhm, K. (2014). Pattern-sensitive time-series anonymization and its application to energy-consumption data. *Open Journal of Information Systems (OJIS)*, *1*(1), 3-22.

[6] 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.

[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] PAJOOHAN, M. R. (2013). PRIVACY PRESERVATION FOR PUBLISHING MEDICAL TIME SERIES: KANONYMIZATION OF NGRAM.

[9] Wik, E. (2019). Anonymously Publishing Univariate Time-Series: With focus on (k, P)-Anonymity.

[10] 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.