Università di Genova
DIBRIS DIPARTIMENTO
DI INFORMATICA, BIOINGEGNERIA,
ROBOTICA E INGEGNERIA DEI SISTEMI

(k,P)-Anonymity

Data Protection & Privacy - a.y. 2019-2020

Giorgio Rossi - S4222099

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. In particular, the spectrum of frequently used “complex” queries on time series covers not only range queries on the attribute values at specified time instants but also pattern similarity queries which treat each sequence more globally

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.

Overview

Why (k,P)-anonymity?

Two levels of anonymization : k, P

NAIVE vs KAPRA approach

NAIVE Algorithm

KAPRA Algorithm

Datasets

Time performance

Future updates

Why (k,P)-anonymity?

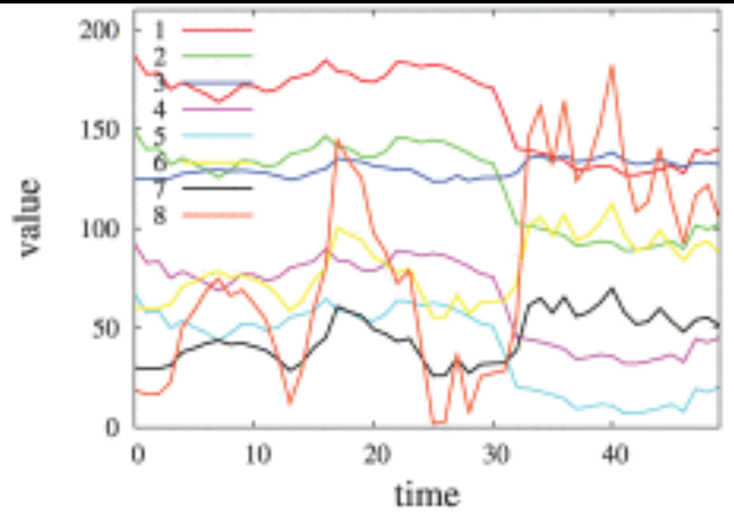
k-anonymity →

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

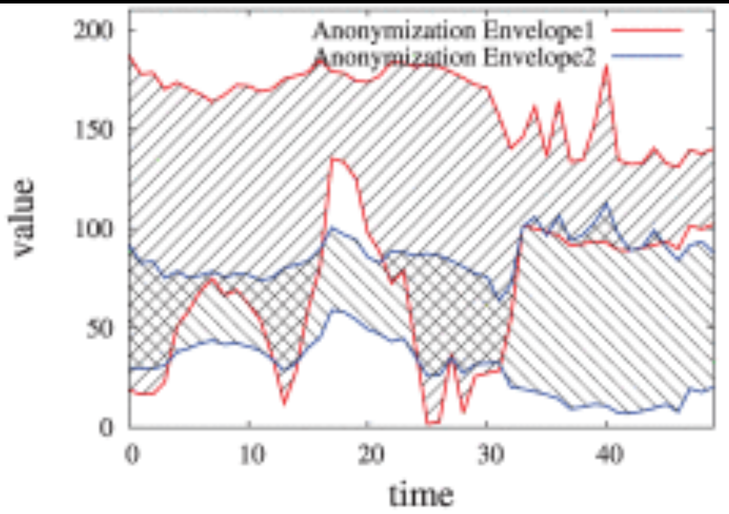
(k,P)-anonymity →

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

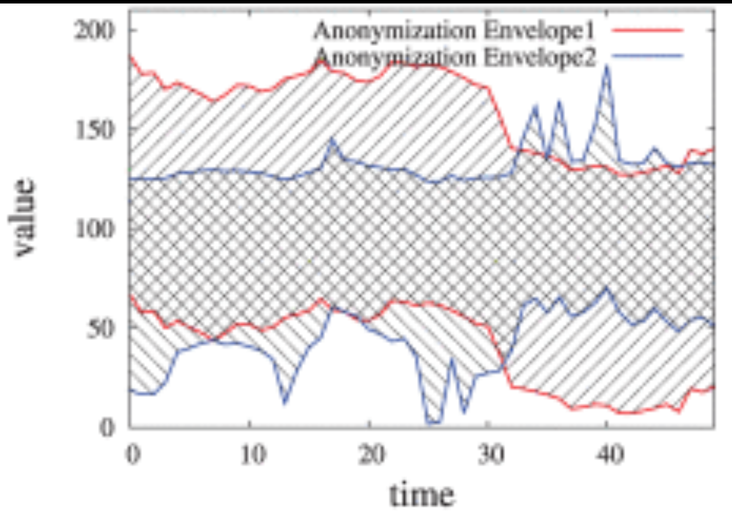
Why (k,P)-anonymity?



(a) Micro data



(b) Generalization result of conventional 4-anonymity. Group 1 contains 1,2,3,8, while group 2 contains 4,5,6,7.



(c) Generalization result of conventional 4-anonymity based on pattern similarity. Group 1 contains 1,2,4,5, while group 2 contains 3,6,7,8.

conventional k-anonymity vs k-anonymity based on pattern similarity

Overview

Why (k,P)-anonymity?

Two levels of anonymization : k, P

NAIVE vs KAPRA approach

NAIVE Algorithm

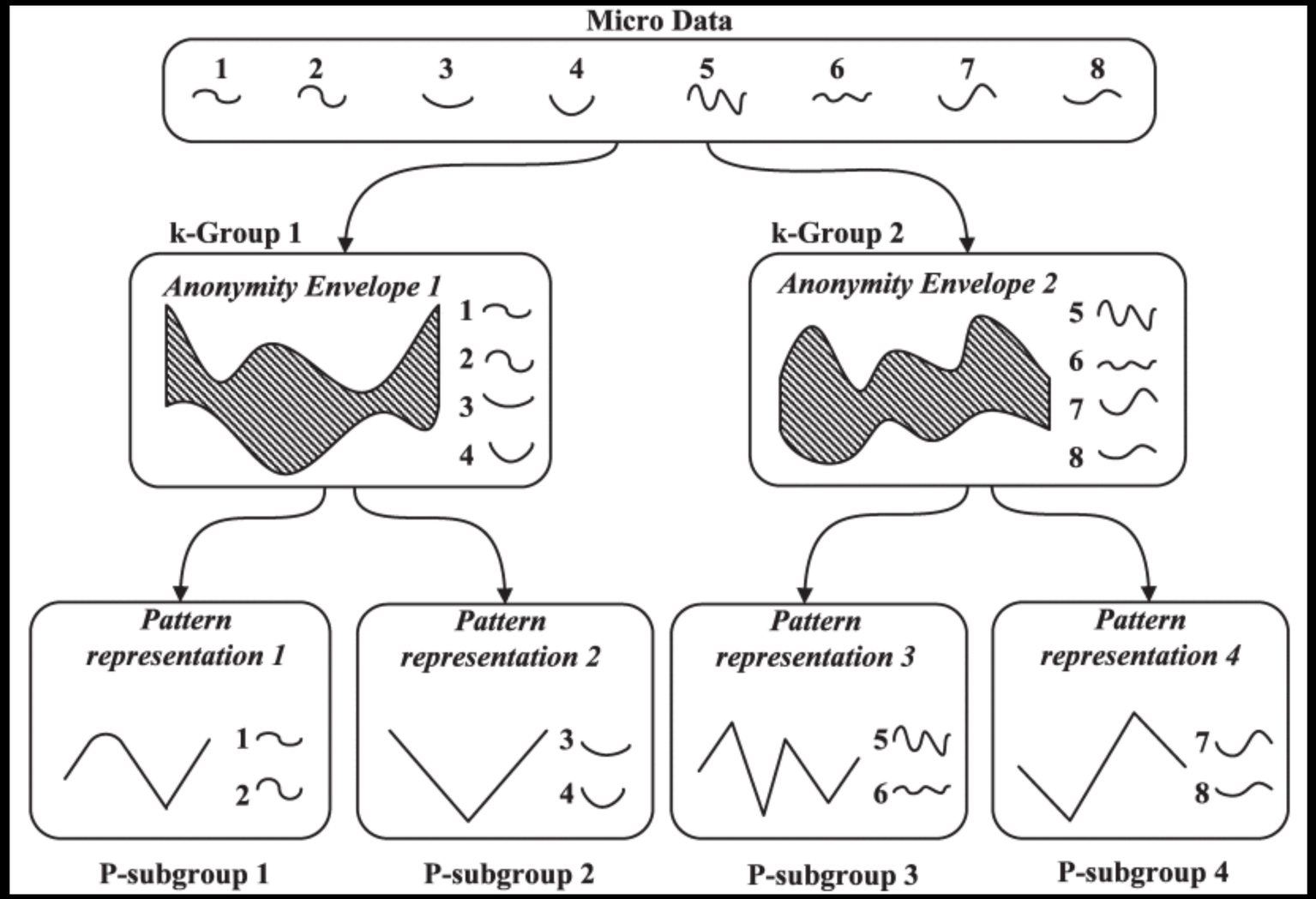
KAPRA Algorithm

Datasets

Time performance

Future updates

Two levels of anonymization: k, P



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)

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

Overview

Why (k,P)-anonymity?

Two levels of anonymization : k, P

NAIVE vs KAPRA approach

NAIVE Algorithm

KAPRA Algorithm

Datasets

Time performance

Future updates

NAIVE vs KAPRA approach

NAIVE Algorithm: top-down clustering (k-anonymity), then create-tree phase (P-anonymity)

KAPRA Algorithm: create-tree phase with entire dataset (P-anonymity) then top-down clustering in atomic group, assembly k-group with group creation phase (k-anonymity)

Overview

Why (k,P)-anonymity?

Two levels of anonymization : k, P

NAIVE vs KAPRA approach

NAIVE Algorithm

KAPRA Algorithm

Datasets

Time performance

Future updates

NAIVE Algorithm

NAIVE Algorithm:

- top-down clustering
  - search two tuples with max uncertainty penalty (group generator)
  - assign other and recursively partition dataset
  - top-down postprocessing
- create-tree phase
  - initialization
  - node splitting
  - postprocessing

NAIVE Algorithm

NAIVE Algorithm:

- top-down clustering
  - search two tuples with max uncertainty penalty
  - assign other and recursively partition dataset
  - top-down postprocessing
- create-tree phase
  - initialization
  - node splitting
  - postprocessing

- weak file path handling

- issue on iteration of NCP(u,v) to find u,v generator
- bug in function that search for max_NCP (always 0)

- not implemented (it drove me crazy)

- issue in node-merge pattern representation

Overview

Why (k,P)-anonymity?

Two levels of anonymization : k, P

NAIVE vs KAPRA approach

NAIVE Algorithm

KAPRA Algorithm

Datasets

Time performance

Future updates

KAPRA Algorithm

KAPRA Algorithm:

- create-tree phase with entire dataset
  - initialization
  - node splitting
- recycle bad-leaves phase
- group formation phase
  - top-down preprocessing
  - group formation
  - group postprocessing

Overview

Why (k,P)-anonymity?

Two levels of anonymization : k, P

NAIVE vs KAPRA approach

NAIVE Algorithm

KAPRA Algorithm

Datasets

Time performance

Future updates

Datasets

Sales transactions - UCI

- ~800 tuples
- 52 weeks (1 y)
- tests

News social feedback - UCI

- ~100000 tuples
- 144 time slices of 20 min (2 d)
- time analysis

Overview

Why (k,P)-anonymity?

Two levels of anonymization : k, P

NAIVE vs KAPRA approach

NAIVE Algorithm

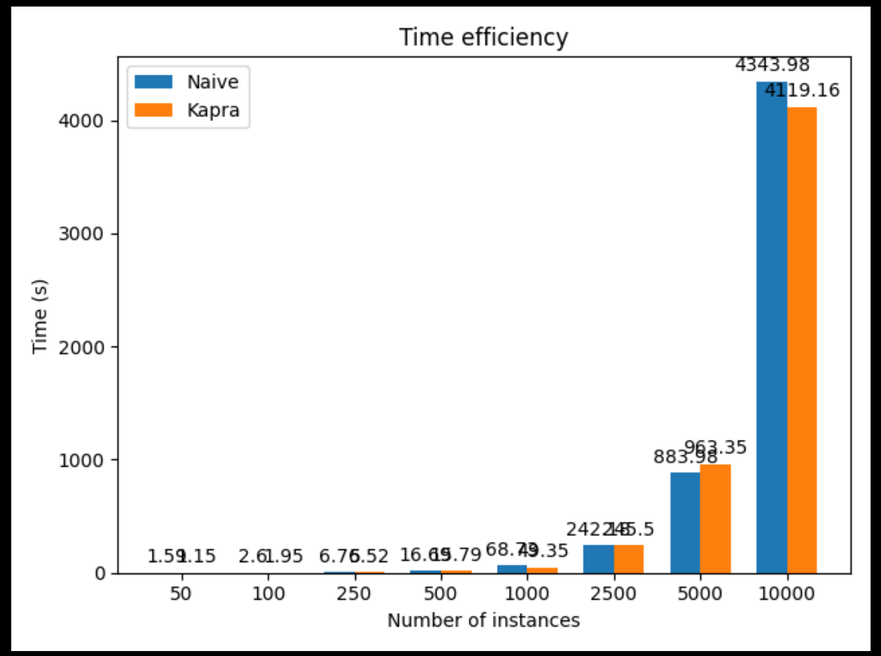
KAPRA Algorithm

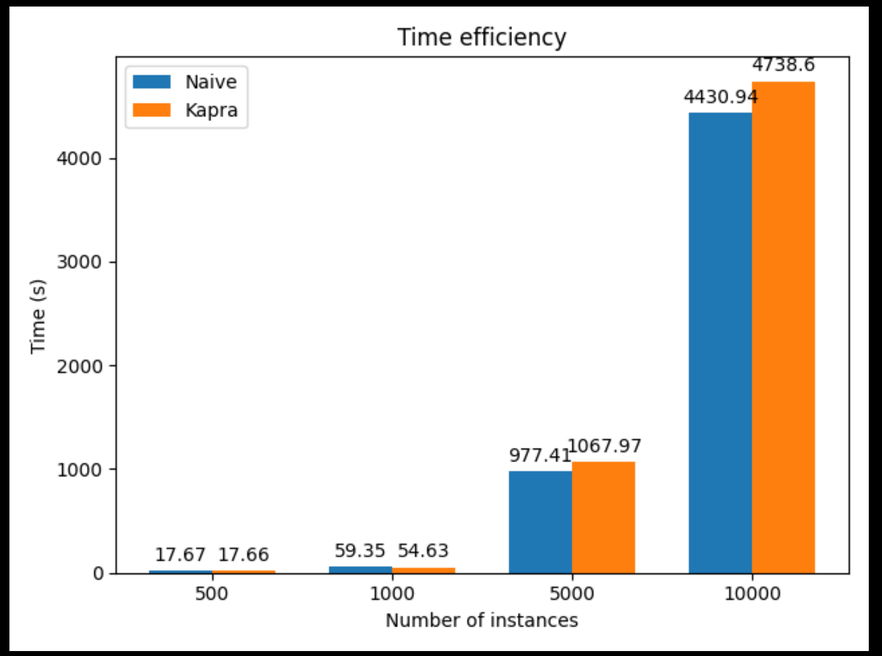
Datasets

Time performance

Future updates

Time performance





Overview

Why (k,P)-anonymity?

Two levels of anonymization : k, P

NAIVE vs KAPRA approach

NAIVE Algorithm

KAPRA Algorithm

Datasets

Time performance

Future updates

Future updates

- Data Structure

Pandas DataFrame to store time series?

dict to store DF, label, level and all anonym. process information?

- Performance optimisation

TODO: reduce function complexity

- Value and Pattern loss report

paper metrics implementation

- Script for tuning hyperparameters

paa, max_level & other parameters tuning

Thank you
for your attention!

References

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.

J. Xu et al., “Utility-Based Anonymization for Privacy Preservation with Less Information Loss,” SIGKDD Explorations, vol. 8, no. 2, pp. 21-30, 2006.