Data anonymization algorithms and data aggregation approaches.

Grigorios Piperagkas , CERTH

CERTH SmartWins Summer School: Day 3

06 July 2023

Thessaloniki

Funded by the European Union

This project has received funding from the European Union's Horizon research and innovation programme under grant agreement No 101078997

SmartWins

Main objectives for data anonymization and aggregation algorithms

- The goal is to deliver high quality and meaningful data even through strong aggregations avoiding too bulky (and unnecessary) information.

- As for data anonymization, strategies can be identified to comply with the General Data Protection Regulation (GDPR), promoting privacy for sensitive data where no consent is required from data owners.

Data anonymization state-of-the-art algorithms

- Three known algorithms for k-anonymization: **Datafly**, **Mondrian and Flash**

- Most common metrics for evaluation of output:
- k-anonymity,
- l-diversity and
- t-closeness.

- Sweeney, Latanya. "**Datafly**: **A system for providing anonymity in medical data**." *Database Security XI*. Springer, Boston, MA, 1998. 356-381.

- LeFevre, Kristen, David J. DeWitt, and Raghu Ramakrishnan. "**Mondrian multidimensional k-anonymity.**" *22nd International conference on data engineering (ICDE'06)*. IEEE, 2006.

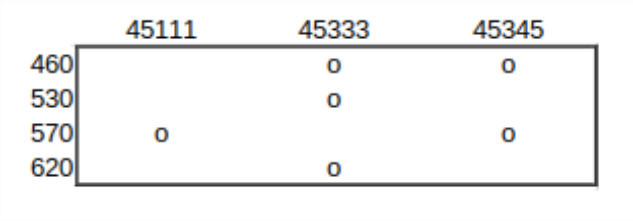
- Kohlmayer, Florian, et al. "**Flash: efficient, stable and optimal k-anonymity.**" *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, 2012.

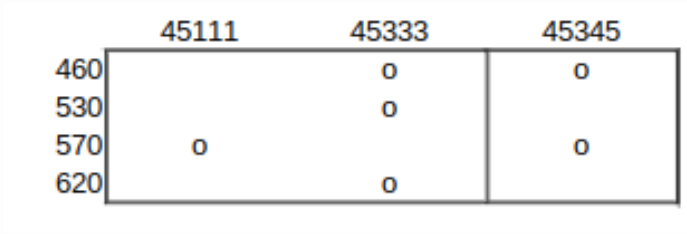
Mondrian k-anonymization algorithm

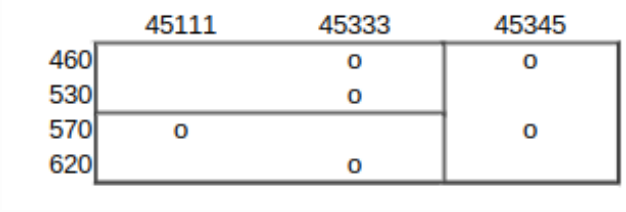
Mondrian is a **greedy partitioning** algorithm

Partitioning is performed recursively to the initial data with a threshold the value k,

meaning that each partitioned table will have **at least k entries,** anonymized, and these are concatenated to the exported table.







Multi-dimensional partitioning process for data e.g. power consumption and zip codes.

Implementation of Mondrian k-anonymization algorithm

Top-down greedy algorithm for strict multidimensional partitioning
anonymize(partition)
if (no allowable multidimensional cut for *partition*)
return *Φ : partition*
else
dim := choose_dimension()
fs := frequency_set(*partition*, *dim*)
splitVal := find_median(*fs*)
lhs := {*t in partition : t.dim < splitVal*}
rhs := {*t in partition : t.dim > splitVal*}
return merge(anonymize(*rhs*), anonymize(*lhs*))

Mondrian k-anonymization algorithm

When partitioning is finished and k-anonymized data groups have been defined, two main processes take place:

-**Suppression**: Each entry is replaced with a non – recognizable value, e.g. “*” to hide information. Useful for outliers.

-**Generalization**: A group of anonymized entries is generalized to contain ranges of values,
e.g. “ages 30-35”, or “energy consumption range 450-670 kWh” instead of accurate values.

Evaluation of Flash anonymization algorithm

Is open source in the **ARX anonymization toolkit**, publicly available: https://arx.deidentifier.org/

- **Depth first greedy algorithm**: searching in a graph of nodes for optimal solution: the k-anonymized data groups

- Applies also efficiently suppression/generalization in numerical or categorical data.

- Provides multiple metrics, data input and output options, e.g. export csv files

- Developed in Java and can be used as an external library: can be integrated with connectors, REST API, data post-processing etc. for use in other components.

Data aggregation approaches

Initial approach for aggregating data:

-Examine the data, i.e. types, format, structure, pre-processing

-Identify requirements, i.e. level, updates, granularity

-Specify aggregation method, i.e. statistical operations etc

-Implement data aggregation, i.e. implement the required algorithm

-Validate results, i.e. ensure aggregates are efficient and meet the requirements.

-Ensure privacy/security, i.e. for sensitive data, ensure GDPR rules by anonymizing or encrypting.

-Provide documentation

Data aggregation approaches

Multiple types of data aggregation

-**Temporal aggregation**: In energy-related senarios, aggregation based on granularity, e.g. hourly aggregates, or daily, monthly averages, sums of energy consumption/production per time period, etc.

-**Spatial aggregation**: can be used in districts, buildings, or area considered as selection for total aggregates. Residential building blocks can be selected, or districts, to evaluate energy consumption/production

-**Aggregation by energy source:** Aggregation per energy production type, e.g. PV, wind, biomass. Aggregation per energy consumption type, e.g. residential, thermal, commercial, mobility.

Data aggregation gains

Data aggregation can be used for functional issues :

-**Balancing supply and demand**: can be used for monitoring energy supply and demand.

-**Optimizing energy use**: Can help operators identify issues of energy consumption inside the districts, inefficiencies etc. and develop strategies.

-**Energy trading:** aggregation of energy data can enable prosumers trading with each other and the grid.

-**Provide useful insights for policy decisions:** provide insights for energy consumption patterns and inform decisions

Data aggregation techniques

Multiple simple and more extensive aggregation methodologies:
- Clustering algorithms, based on similarities in usage patterns

- Time-series analysis, for identification of patterns and trends

- Fourier transform, for analyzing frequency components of energy data

- ML algorithms, e.g. Principal Component Analysis, Artificial Neural Networks, to simplify, model energy data and make predictions.

Most common approaches:
-**Descriptive statistics**: mean, median, StD, ranges, to describe characteristics
-**Correlation analysis:** analyzing relationship between variables
-**Regression:** statistical models to predict energy consumption/production
-**Hypothesis testing:** testing hypothesis for finding significant difference in examined data.

Example results for k-anonymization algorithms

Name	Age	Gender	consumption
Alice	*21*	*F*	*600*
Bob	*40*	*M*	*720*
Charlie	*24*	*M*	*450*
David	*21*	*M*	*380*
Alice	*45*	*F*	*670*
Jo	*30*	*M*	*445*
Albert	*25*	*M*	*562*
John	*30*	*M*	*455*
Alice	*30*	*F*	*930*
Jonathan	*21*	*M*	*790*
Bob	*42*	*M*	*560*
Elle	*41*	*F*	*445*
Alice	*27*	*F*	*602*
Eve	*45*	*F*	*430*

Example results for k-anonymization algorithms, k=2

Name	Age	Gender	consumption
*	*	*	*
2-3	21-24	1	380-450
2-3	21-24	1	380-450
4-6	30	1	445-455
4-6	30	1	445-455
8-9	41-45	0	430-445
8-9	41-45	0	430-445
*	*	*	*
0-5	25-27	0-1	562-602
0-5	25-27	0-1	562-602
0-7	21-30	0-1	790-930
0-7	21-30	0-1	790-930
0-1	40-45	0-1	670-720
0-1	40-45	0-1	670-720

Example results for k-anonymization algorithms, k=3

Name	Age	Gender	consumption
1-2	*21-31.5*	*1*	*380-470*
2-3	*21-31.5*	*1*	*380-470*
1-2	*31.5-42*	*1*	*470-560*
*	*	*	*
4-6	*30*	*1*	*445*
4-6	*30*	*1*	*455*
6-8	*41*	*0*	*445*
0	*21-24*	*0*	*582-602*
5	*24-27*	*1*	*562-582*
0	*24-27*	*0*	*582-602*
*	*	*	*
*	*	*	*
*	*	*	*
*	*	*	*

Thank you

Questions?

Project Partners

ktu kaunas university of technology 1922

CERTH CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

POLITECNICO MILANO 1863

Contecht

INNOTROPE

Funded by the European Union

This project has received funding from the European Union's Horizon research and innovation programme under grant agreement No 101078997

SmartWins