Data Protection & Privacy

a.a. 2021/2022

CAHD :

Working on anonymization of Transaction Data

Simone Cella

Introduction:

This project aims to build an implementation of CAHD (Correlation Aware Anonymization of High-dimensional Data).

CAHD is really helpful in the anonymization process of Transaction Data, but some steps are needed to prepare the Dataset and then work with that.

Step By Step

In order to achieve our result, the anonymized groups from the working dataset, i created 4 python’s class:

- makeDataset
- Dataset
- Algorithm
- KL Divergence

Transaction Data

	Wine	Strawberries	Meat	Cream	Pregnancy Test	Viagra
Bob	X		X			X
David	X		X			
Claire		X		X	X	
Andrea		X	X			
Ellen	X		X	X		

Fig : example of transaction data.
(As we can see the table is sparse)

In order to start the project we need to transform the dataset in a form of Transaction Data

Transaction Data are used to register customer's transactions and to extract information (like the correlation between transactions)

All the transactions are important but we will discover that some are more important than other.

class : makeDataset:

This class operates in this way:

Take as input the dataset path, the file txt containing the items, and a delimiter.
The delimiter is used to decide where the algorithm needs to stop in order to create the dataset.

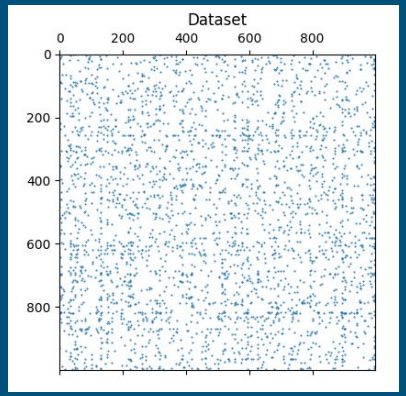
After that it creates the Dataframe from the csv file given before, and creates all the instances that we need from the Dataset :

-items, transactions, transaction matrix...

class : Dataset

The dataset class, instead, from this working file, it creates the Band Matrix* of the dataset and after that will generate *n* Sensitive Items, it creates a part of the Data Frame containing only the Sensitive transactions.

* first we divided QIDs from SIs and with the matrix of QIDs we perform a permutation of rows and columns in order to have close transactions as similar as possible.



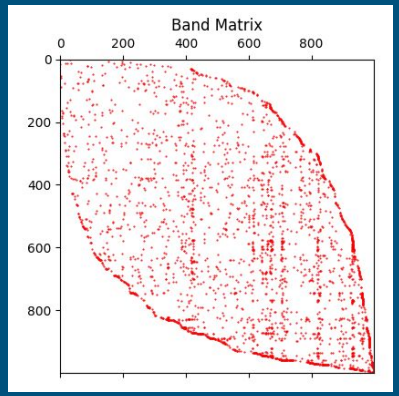


Fig: Realization of a band matrix from the BMS2 dataset
This is realized using the Reverse Cuthill-McKee (RCM) algorithm

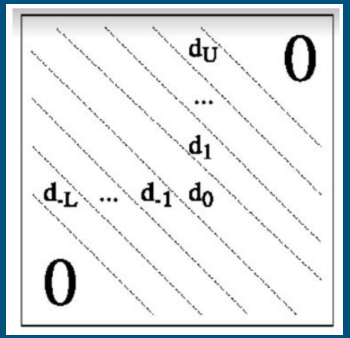


Fig: example of a generic Band Matrix

Our objective in making the Band Matrix is to minimize the total bandwidth

$$B = U + L + 1$$

CAHD (classs : Algorithm)

One of the proposed solutions, in order to make anonymized group from a dataset of transaction data, is the ***Correlation-Aware Anonymization of High-Dimensional Data*** **(CAHD).**

This approach needs to satisfy two important requirements:

- ***Privacy Requirements:***

  a privacy-preserving transformation of the transaction set T has privacy degree p if the probability of associating any transaction t in T with a particularly sensitive item s in S doesn't exceed 1/p.

- ***Utility Requirements:***

  In order to preserve the privacy of transaction data, a certain amount of information loss is inherent, in order to minimize the reconstruction error it is necessary to group together transactions with similar QIDs.
  (See band matrix before)

CAHD Step1 to Step2:

Step 1:

	Wine	Strawberries	Meat	Cream	Pregnancy Test	Viagra
Bob	X		X			X
David	X		X			
Claire		X		X	X	
Andrea		X	X			
Ellen	X		X	X		

(a) Original Data

	Wine	Meat	Cream	Strawberries	Pregnancy Test	Viagra
Bob	X	X				X
David	X	X				
Ellen	X	X	X			
Andrea		X		X		
Claire			X	X	X	

(b) Re-organized Data

	Wine	Meat	Cream	Strawberries	Sensitive Items
Bob	X	X			Viagra: 1
David	X	X			
Ellen	X	X	X		
Andrea		X		X	Pregnancy Test: 1
Claire			X	X	

(c) Published Groups

The first step, as already mentioned is to prepare the working dataset in a way that the reconstruction error is minimized. One of the most powerful way is to rearrange the transactions as shown in the fig above.

Step 2:

	Wine	Strawberries	Meat	Cream	Pregnancy Test	Viagra
Bob	X		X			X
David	X		X			
Claire		X		X	X	
Andrea		X	X			
Ellen	X		X	X		

(a) Original Data

	Wine	Meat	Cream	Strawberries	Pregnancy Test	Viagra
Bob	X	X				X
David	X	X				
Ellen	X	X	X			
Andrea		X		X		
Claire			X	X	X	

(b) Re-organized Data

	Wine	Meat	Cream	Strawberries	Sensitive Items
Bob	X	X			Viagra: 1
David	X	X			
Ellen	X	X	X		
Andrea		X		X	Pregnancy Test: 1
Claire			X	X	

(c) Published Groups

The second step is to create, starting from the Sensitive Transactions, the whole group of transactions to be added to it, and then making the anonymized groups.

Step 2 - Explained

CAHD Group Formation Heuristic

Input: transaction set T, privacy degree p

1. initialize histogram H for each sensitive item $s \in S$
2. $remaining = |T|$
3. **while** $(\exists t \in T | t \ is \ sensitive)$ **do**
4. t = next sensitive transaction in T
5. $CL(t)$ = non-conflicting αp pred. and αp succ. of t
6. $G = \{t\} \cup p-1$ trans. in $CL(t)$ with closest QID to t
7. update H for each sensitive item in G
8. **if** $(\nexists s | H[s] \cdot p > remaining)$
9. $remaining = remaining - |G|$
10. **else**
11. roll back G and continue
12. **end while**
13. output remaining transactions as a single group

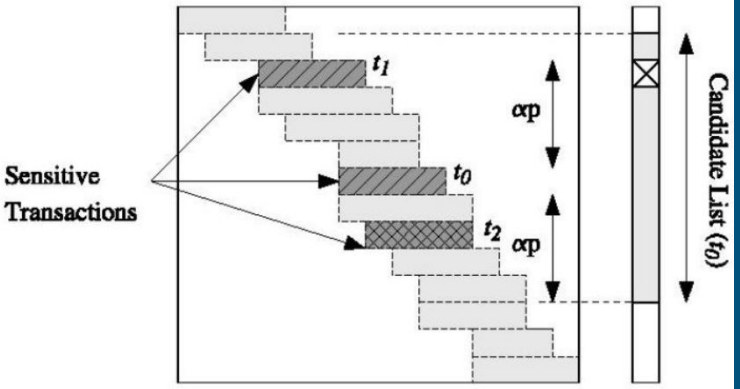


Fig. 7. Group Formation Heuristic

From a sensitive transaction, one list of candidate (CL) need to be formed, with &p transaction that preceding or following it (and obviously are not in conflict with this transaction). So, after has formed, this Candidate List group, the algorithm will go to discover the most similar in the group, in order to preserve the information loss.
All the transaction are deleted in the group of transaction T, and the process continue until there is no more sensitive transaction in the group.

class : KL Divergence

By implementing the calculus of the distribution of probability (***pdf***) of an item ***s*** in a fitted space of item in ***G***, it's one of the best way to evaluate the information loss.

The value of pdf of an certain sensibile ***s*** item s in ***c*** (one cell) is:

$$Act_C^s = \frac{Occurrences\ of\ s\ in\ C}{Total\ Occurrences\ of\ s\ in\ T}$$

The other useful value, Est_c^s the numerator, instead of occurrences of ***s*** in ***C*** will be the total possible combination of the groups that intercept the cell ***C***.

Finally, we can compute the Kl Divergence :

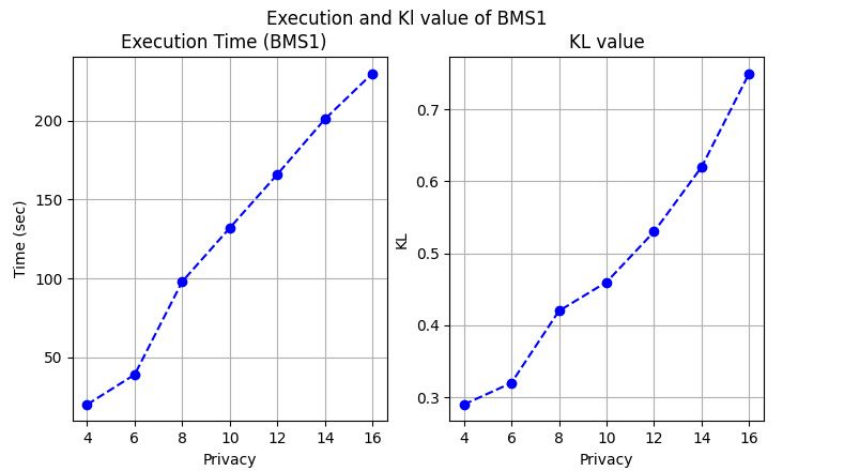
$$KL\_Divergence(Act, Est) = \sum_{\forall cell\ C} Act_C^s \log \frac{Act_C^s}{Est_C^s}$$

If Act == Est → Kl Divergence = 0

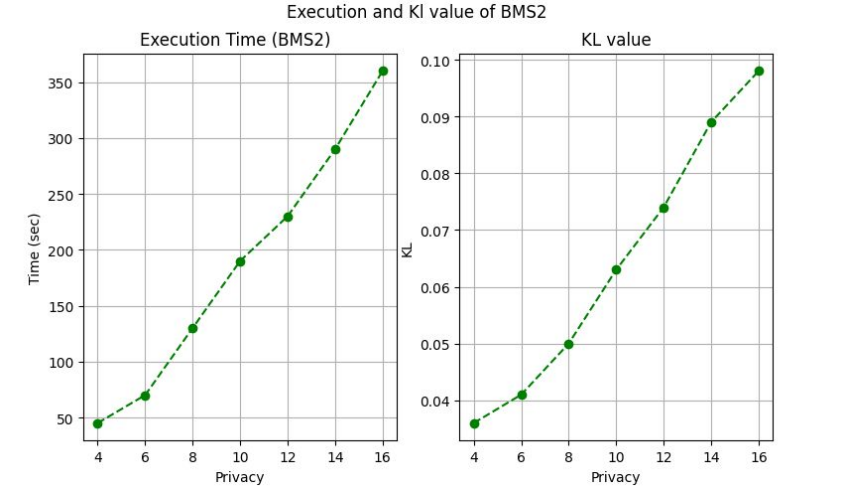
The objective of this project is to minimize the reconstructor error, measured by Kl Divergence.

Some test:

All the tests were done with this indices:
- 1000 items
- 10 sensible items
- alpha = 3



BMS1 Dataset



BMS2 Dataset

I appreciate your attention!

Simone Cella