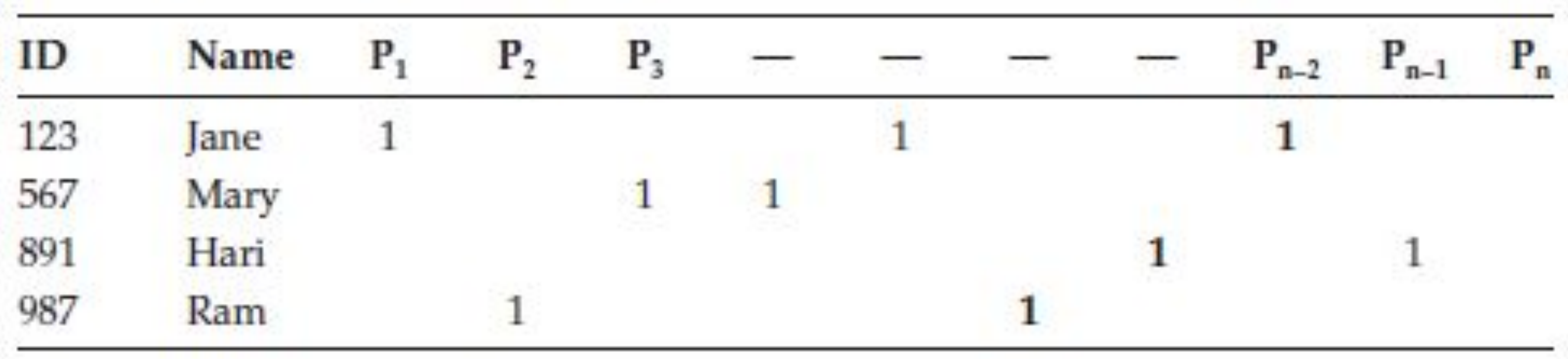
Anonymizing Transaction Data: Correlation-aware Anonymization of High-dimensional Data (CAHD)

UNIVERSITÀ DEGLI STUDI DI GENOVA | Dibris

Davide Caputo
davide.caputo@dibris.unige.it

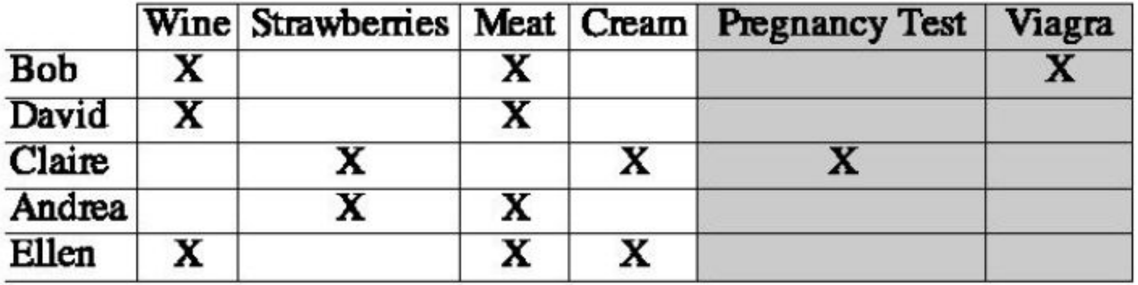
Example of Transaction Data



ID	Name	P_1	P_2	P_3	—	—	—	—	P_{n-2}	P_{n-1}	P_n
123	Jane	1				1			1		
567	Mary			1	1						
891	Hari							1		1	
987	Ram		1				1				

Characteristics of Transaction Data

- The table is sparse, very few cells have entries in this high-dimensional space
- Mined to extract *associations* or *correlation* among transactions
- Exists Sensitive Items and Quasi-Identifier Items



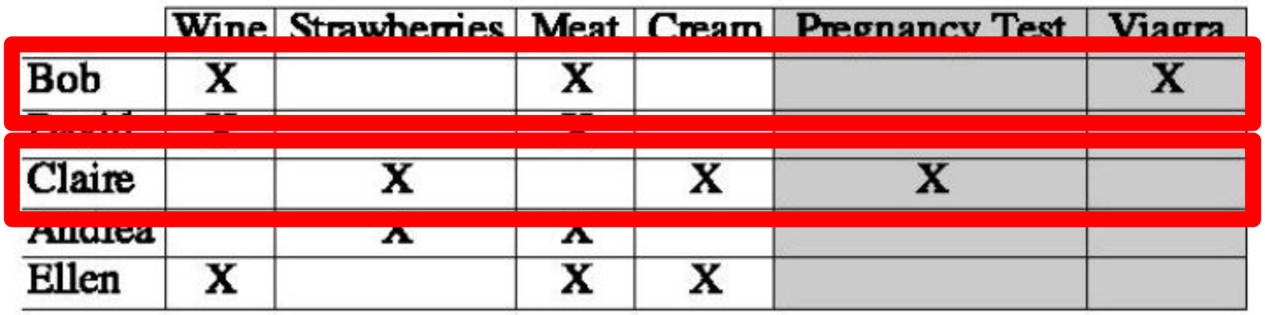
	Wine	Strawberries	Meat	Cream	Pregnancy Test	Viagra
Bob	X		X			X
David	X		X			
Claire		X		X	X	
Andrea		X	X			
Ellen	X		X	X		

(a) Original Data

Characteristics of Transaction Data (2)

- We define a privacy breach if we are able to associate a sensitive product to a certain individual (we must prevent this association)

- There are few sensitive transactions that are classified as sensitive



	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

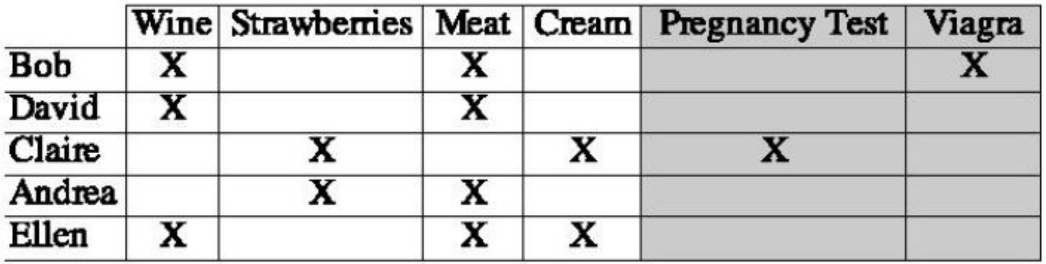
Goal and Definitions

- Our objective is to anonymized data consisting of a set of transaction $T = \{t1, t2, \ldots, tn\}$, $n= |T|$.
- Each transaction $t \in T$ contains items from an item set $I = \{i1, \ldots, id\}$, $d = |I|$.
- Among the set of items I, some are privacy-sensitive (such as pregnancy test or viagra)
- **Privacy-sensitive**: The set $S \in I$ of items that represent a privacy threat if associated to a certain transaction, constitutes the sensitive items set, $S = \{s1, \ldots, sn\}$, $m = |S|$.
- The rest of items in I are non sensitive, and we denote these items by **Quasi-identifier (QID)** items

Data Representation

- We represent the data as a binary matrix A with n rows and d columns

$$A[i][j] = \begin{cases} 1, & i_j \in t_i \\ 0, & i_j \notin t_i \end{cases}$$



	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

$$A = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}$$

Requirements of Anonymized Transaction Data

- The Anonymized Transaction data should satisfy two requirements:
    - Privacy Requirements
    - Utility Requirements

Privacy Requirements

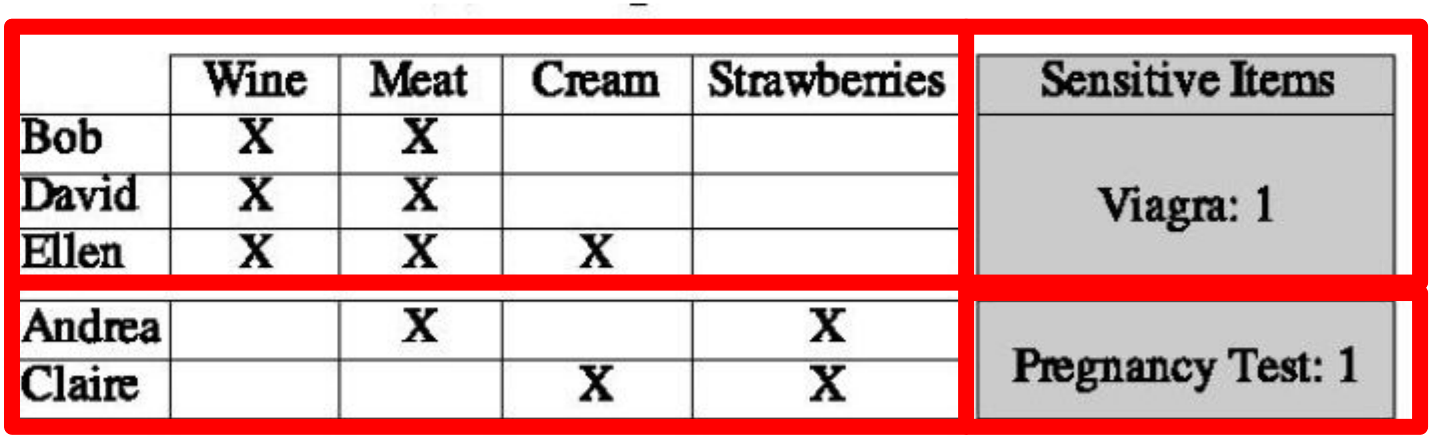
- Privacy: A privacy-preserving transformation of transaction set T has *privacy degree* **p** if the probability of associating any transaction t ∈ T with a particular sensitive item s ∈ S does not exceed ***1/p***.

- We enforce the privacy requirement by partitioning the set T into disjoint sets of transactions, which we refer to as anonymized groups

- For each group G, we publish the exact QID items, together with a summary of the frequencies of sensitive items contained in G.

Example

The probability of associating any transaction in G to sensitive item is 1/2.



	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

Privacy Requirements (2)

- In general, let $f1^G$, … $f2^G$ be the number of occurrences for sensitive items s1, …, sm in group G. Then group G offers privacy degree

$$p^G = \min_{i=1\ldots m} |G| / f_i$$

- The privacy degree of an entire partitioning P of T is

$$p^{\mathcal{P}} = \min_{G \in \mathcal{P}} p^G$$

Utility Requirements

- In order to preserve privacy of transactional data, a certain amount of information loss is inherent.

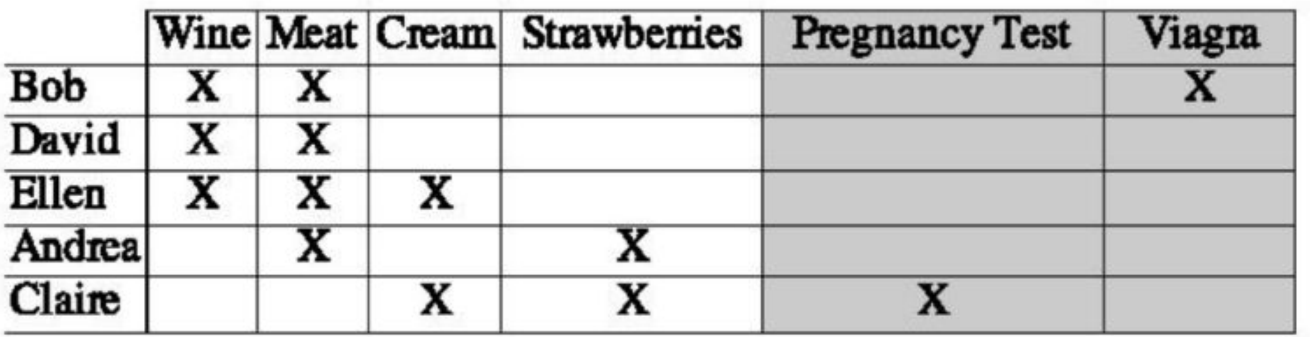
- But the data should maintain a reasonable degree of utility

- Transactional data is mainly utilized to derive certain patterns, such as consumer purchasing habits.

Utility Requirements (2)

- In order to minimize the reconstruction error, it is necessary to group together transactions with similar QID

- The data (i.e. the matrix A) are organized in a **band matrix**, so that consecutive rows are likely to share a large number of common items



	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	

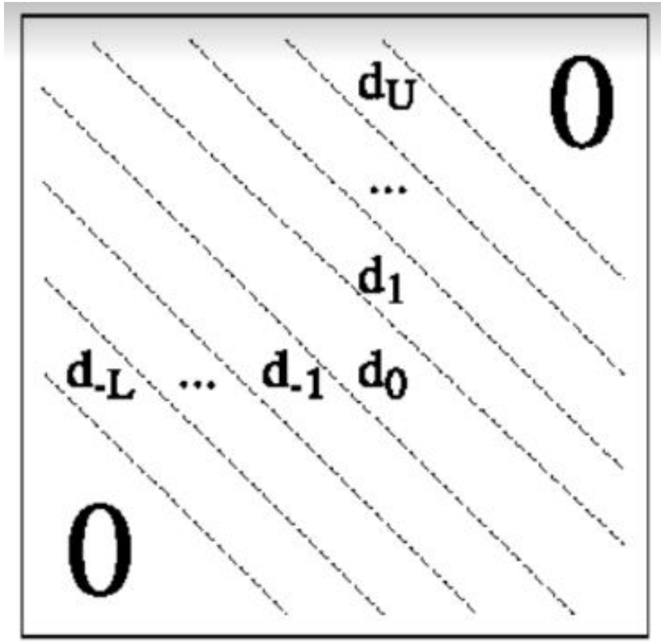
(b) Re-organized Data

Utility Requirements (3)

- A **band matrix** has **0** on all elements of the matrix, except for the main diagonal d_0, a number of *U* upper diagonals (d_1,...,d_U) and *L* lower diagonals (d_{-1},...,d_{-L})

- *U* represents the *upper bandwidth* of the matrix and *L* the *lower bandwidth*

- Our objective is to minimize the total bandwidth **B= U+L+1**



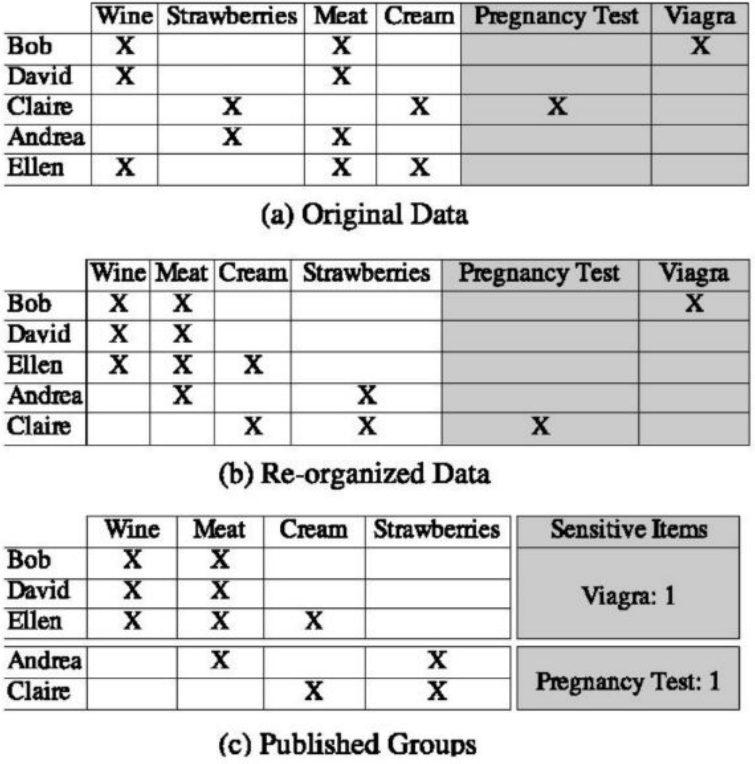
Correlation-aware Anonymization of High-dimensional Data (CAHD)

The Correlation-aware Anonymization of High-dimensional Data (CAHD) algorithm is based on two steps:

1. Create Band Matrix using Reverse Cuthill-McKee Algorithm (RCM) to fulfill utility requirement (scipy.sparse.csgraph.reverse_cuthill_mckee)

2. Create Anonymized Groups to fulfill privacy requirement

Example of CAHD



	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	

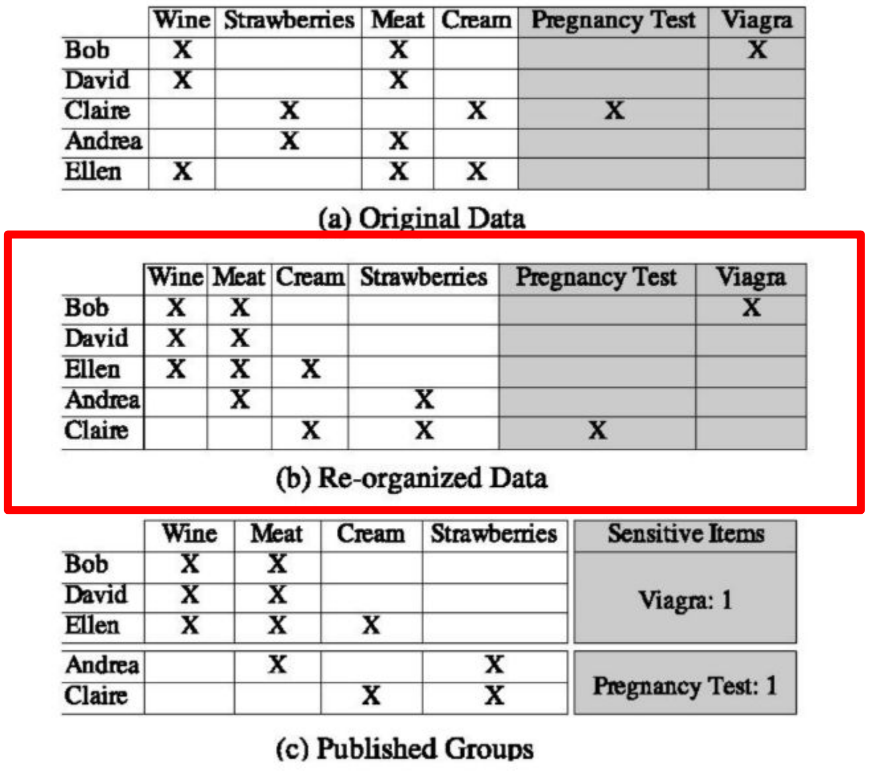
(b) Re-organized Data

	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

Example of CAHD

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	

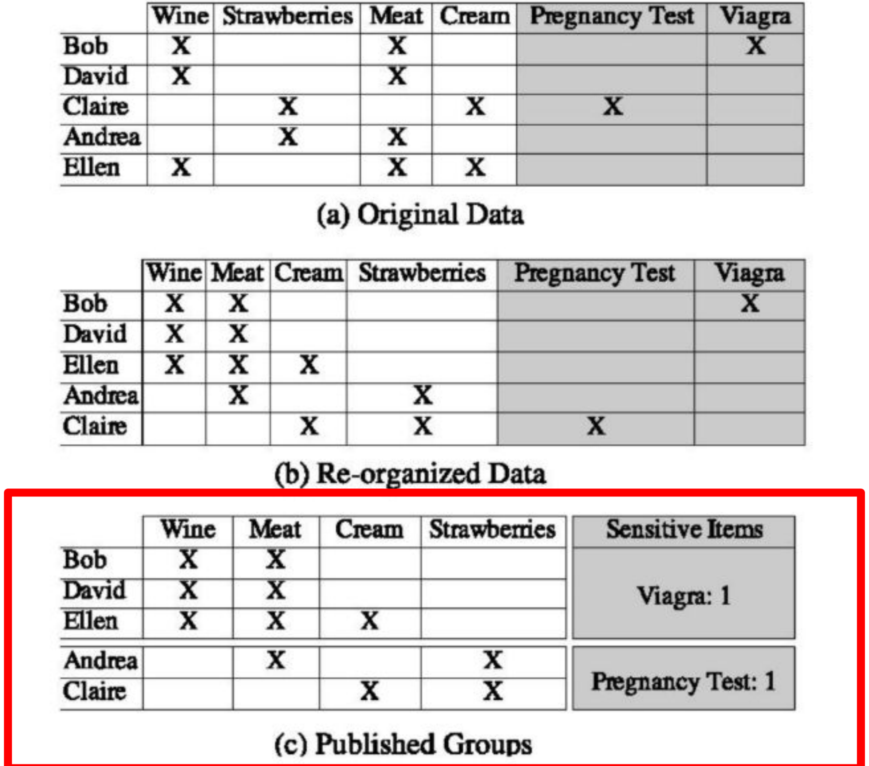
(b) Re-organized Data

	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

Example of CAHD

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

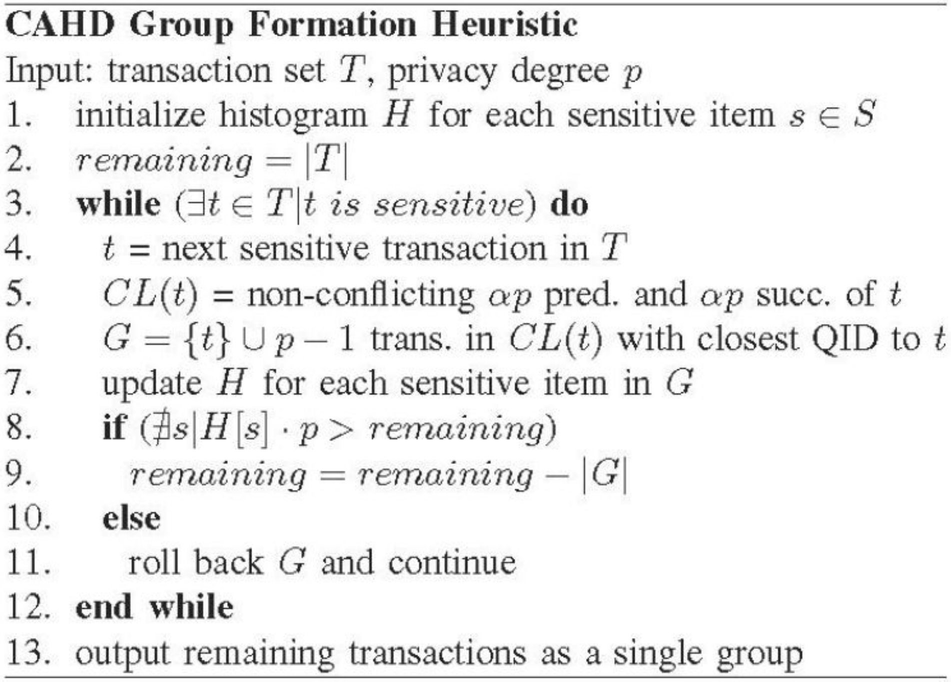
	Wine	Meat	Cream	Strawberries	Pregnancy Test	Viagra
Bob	X	X				X
David	X	X				
Ellen	X	X	X			
Andrea		X		X		
Claire			X	X	X	

(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

Step 2 - Create Anonymized Groups



CAHD Group Formation Heuristic

```
Input: transaction set $T$, privacy degree $p$
initialize histogram $H$ for each sensitive item $s \in S$
$remaining = |T|$
while $(\exists t \in T | t \ is \ sensitive)$ do
    $t$ = next sensitive transaction in $T$
    $CL(t)$ = non-conflicting $\alpha p$ pred. and $\alpha p$ succ. of $t$
    $G = \{t\} \cup p - 1$ trans. in $CL(t)$ with closest QID to $t$
    update $H$ for each sensitive item in $G$
    if $(\nexists s | H[s] \cdot p > remaining)$
        $remaining = remaining - |G|$
    else
        roll back $G$ and continue
end while
output remaining transactions as a single group
```

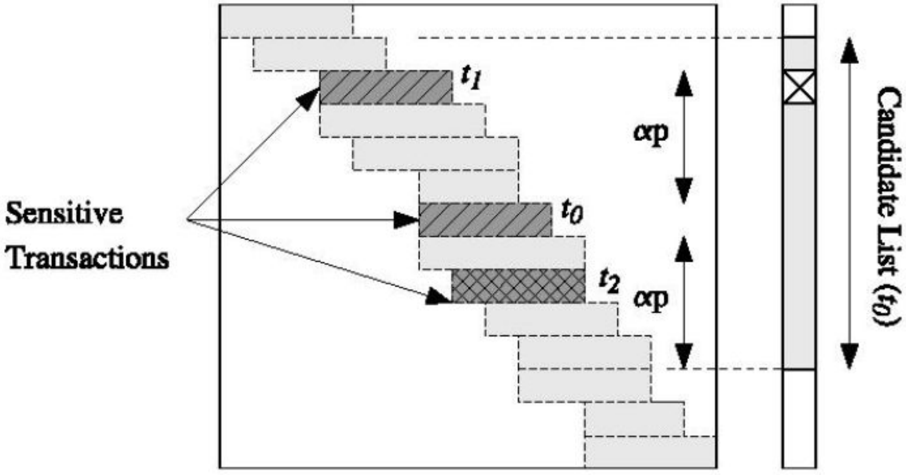


Fig. 7. Group Formation Heuristic

Usually α=3

References

1. On the Anonymization of Sparse High-Dimensional Data (https://ieeexplore.ieee.org/document/4497480)