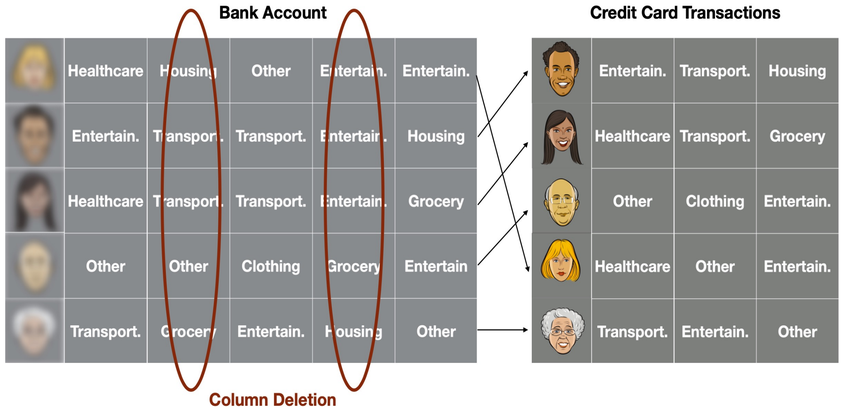
Database Matching Under Column Deletions

Serhat Bakirtas, Elza Erkip
serhat.bakirtas@nyu.edu, elza@nyu.edu
NYU Tandon School of Engineering

I. Motivation

- ❖ Personal data published or sold after anonymization
- ❖ Anonymization is not enough!
  - Correlated Data → De-anonymization
- ❖ Motivated by time-indexed databases
  - Synchronization errors while sampling
  - Column deletion
  - Deletion locations are unknown!



II. Objectives

What are the sufficient conditions for successful de-anonymization?

How does side information on the deletion locations help?

Can we extract this side information from an already-matched batch of rows?

How large does this batch have to be?

III. Problem Formulation

$\mathcal{C}^{(1)}$ Attribute Vector

1	$X_{1,1}$	$X_{1,2}$	• •	$X_{1,n-1}$	$X_{1,n}$
2	$X_{2,1}$	$X_{2,2}$	• •	$X_{2,n-1}$	$X_{2,n}$
•	•	•	• •	•	•
•	•	•	• •	•	•
m	$X_{m,1}$	$X_{m,2}$	• •	$X_{m,n-1}$	$X_{m,n}$

$(\mathcal{C}^{(2)}, \Theta)$ User ID Attribute Vector

$\Theta^{-1}(1)$	$X_{\Theta^{-1}(1),1}$	$X_{\Theta^{-1}(1),3}$	• •	$X_{\Theta^{-1}(1),n-1}$
$\Theta^{-1}(2)$	$X_{\Theta^{-1}(2),1}$	$X_{\Theta^{-1}(2),3}$	• •	$X_{\Theta^{-1}(2),n-1}$
•	•	•	• •	•
•	•	•	• •	•
$\Theta^{-1}(m)$	$X_{\Theta^{-1}(m),1}$	$X_{\Theta^{-1}(m),3}$	• •	$X_{\Theta^{-1}(m),n-1}$

m: # of users (rows)
n: # of attributes (columns)
δ: column deletion probability
α: deletion detection probability

Database Growth Rate

$$R = \lim_{n\to\infty} \frac{1}{n} \log_2 m$$

IV. Proposed Matching Scheme

Y	$X_{\Theta^{-1}(1),1}$	$X_{\Theta^{-1}(1),3}$	• •	$X_{\Theta^{-1}(1),n-1}$

1	$X_{1,1}$	$X_{1,2}$	• •	$X_{1,n-1}$	$X_{1,n}$
2	$X_{2,1}$	$X_{2,2}$	• •	$X_{2,n-1}$	$X_{2,n}$
•	•	•	• •	•	•
•	•	•	• •	•	•
m	$X_{m,1}$	$X_{m,2}$	• •	$X_{m,n-1}$	$X_{m,n}$

Discarded

X	$X_{j,1}$	$X_{j,2}$	$X_{j,3}$	• •	$X_{j,n-1}$

- ➢ Discard the detected deleted columns from $\mathcal{C}^{(1)}$
- ➢ Match **Y** with **X** if
  1. **X** is ε-typical with respect to p_X
  2. **Y** is a subsequence of X
  3. There is no such other row **X**

V. Achievable Database Growth Rate

Theorem

Given a column deletion probability $\delta < 1 - \frac{1}{|\mathfrak{X}|}$ and a deletion detection probability α, any database growth rate

$$R < \left[(1-\alpha\delta)\left(H(X) - H_b\left(\frac{1-\delta}{1-\alpha\delta}\right)\right) - (1-\alpha)\delta\log(|\mathfrak{X}|-1)\right]^+$$

is achievable, where H,H_b and $[.]^+$ denote the entropy, the binary entropy, and the positive part functions respectively.

VI. Deletion Detection

Given a batch of B pairs of correctly–matched rows.

$$\mathcal{D}^{(1)} = \begin{bmatrix} 0 & 1 & \mathbf{0} & 1 & 1 \\ 1 & 0 & \mathbf{0} & 1 & 1 \end{bmatrix} \qquad \mathcal{D}^{(2)} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

Detect the j-th column $\boldsymbol{D_j}$ of $\mathcal{D}^{(1)}$ to be deleted if

1. $\boldsymbol{D_j}$ is ε-typical with respect to p_X
2. $\boldsymbol{D_j}$ is a column of $\mathcal{D}^{(2)}$

Theorem

Let $\mathcal{D}^{(1)}, \mathcal{D}^{(2)}$ be a batch of correctly-matched B rows of the unlabeled database $\mathcal{C}^{(1)}$, and the corresponding column deleted database $\mathcal{C}^{(2)}$. Then

$$P(g(\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, j) = 1 | j \in I_D) \geq 1 - \epsilon - n2^{-B(H(X)-\epsilon)}(1-\delta)$$

where I_D is the set of deleted column indices.

VII. Observations

➢ Achievable Database Growth Rate

- Decreases with δ
- Increases with α and $H(X)$

➢ Deletion Detection Probability

- Decreases with δ
- Increases with B and $H(X)$

VIII. Conclusion

- ➢ Deletion detection helps!
  - No detection → Deletion Channel
  - Full detection → Erasure Channel
- ➢ Deletion detection can be performed given seeds.
- ➢ A seed size $B = \omega(\log n) = \omega(\log\log m)$ is enough for full detection.

➢ Ongoing work

- Batchwise Matching
- Converse Results