Seeded Database Matching Under Synchronization Errors

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

I. Introduction

❖ **Motivation:** Personal data published after anonymization
- **Practical Works :**Correlated Public Data → De-anonymization

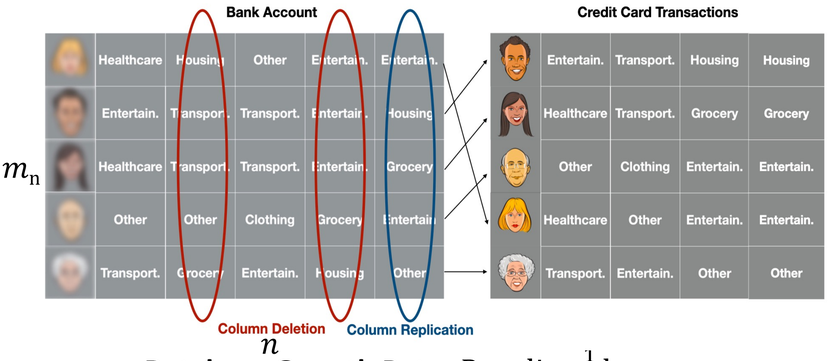
❖ **This Work:** A general rigorous formulation covering various forms of distortion such as noise and synchronization errors

II. System Model

❖ **System Model:**
- **Unlabeled Database:** Random $m \times n$ matrix $\boldsymbol{D}^{(1)}$ generated according to $\boldsymbol{p_X}$
- **Labeling Function:** Uniform permutation Θ_n.
- **Synchronization Errors:** Random column replication and deletion pattern S^n, $\boldsymbol{p_S}$.
- **Noise:** Independent, memoryless. $\boldsymbol{p_{Y|X}}$
- **Labeled Database:** Matrix-permutation pair $(\boldsymbol{D}^{(2)}, \Theta_n)$

$$D^{(2)}_{i,j} = \begin{cases} E, & \text{if } S_j = 0 \\ Y_i^{S_j} & \text{if } S_j \geq 1 \end{cases} \quad \forall i \in [m_n],\ \forall j \in [n]$$

$$\Pr\left(Y_i^{S_j} = y^{S_j} \middle| D^{(1)}_{\Theta_n^{-1}(i),j}\right) = \prod_{l=1}^{S_j} p_{Y|X}\left(y_l \middle| D^{(1)}_{\Theta_n^{-1}(i),j}\right)$$



- **Database Growth Rate:** $R = \lim_{n\to\infty} \frac{1}{n} \log_2 m_n$
- **Seeds**: A batch $(\boldsymbol{G}^{(1)}, \boldsymbol{G}^{(2)})$ of $\Lambda_n = \Theta(n^d)$ already-matched row pairs. (d: Seed Order)

❖ **Successful Matching:** Correct estimation of the labeling function Θ by the attacker, given $(\boldsymbol{D}^{(1)}, \boldsymbol{D}^{(2)}, \boldsymbol{G}^{(1)}, \boldsymbol{G}^{(2)})$

❖ **Matching Capacity:**

$$C(d) = \sup\{R: R \textit{ is achievable given } \Theta(n^d) \textit{ seeds}\}$$

III. Objectives

What are the fundamentals of database matching, including the sufficient and the necessary conditions for successful matching?

How do we perform successful matching under noise and synchronization errors?

IV. Main Result: Matching Capacity

Main Result: Matching Capacity

Consider a database distribution p_X, a column repetition distribution p_S and a noise distribution $p_{Y|X}$. Then, for any seed order $d \geq 1$, the matching capacity is

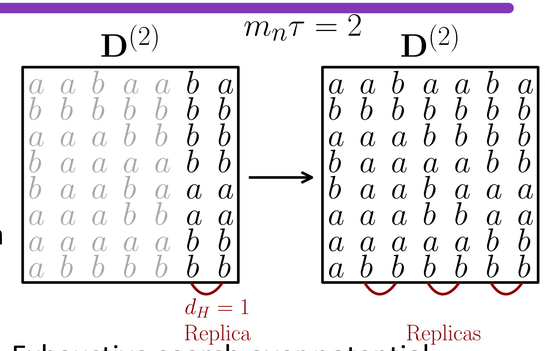
$$C(d) = I(X; Y^S, S)$$

where $S \sim p_S$ and $Y^S = Y_1, \ldots, Y_S$ such that

$$\Pr(Y^S = y_1, \ldots, y_S | X = x) = \prod_{i=1}^{S} p_{Y|X}(y_i|x)$$

V. Achievability-I: Replica and Deletion Detection

❖ **Replica Detection:** Binary Hypothesis Testing on the Hamming distances between consecutive column pairs against a threshold τ



❖ **Deletion Detection:** Exhaustive search over potential deletion indices given seeds $(\boldsymbol{G}^{(1)}, \boldsymbol{G}^{(2)})$.

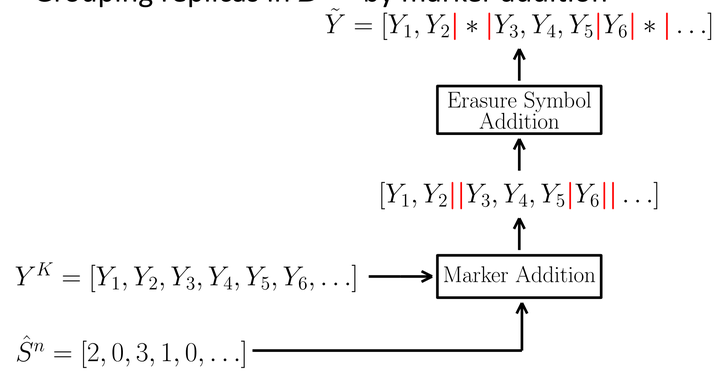
Theorem: Repetition Detection

Let S^n and $\hat{S}^n$ be the repetition pattern and its estimate. Given the seed size $\Lambda_n = \Theta(n)$,

$$\Pr(\hat{S}^n = S^n) \to 1 \text{ as } n \to \infty$$

VI. Achievability-II: Matching Scheme

Matching Scheme :
1. Use the described replica and deletion detection algorithms
2. Convert the problem into a channel decoding problem by
   - Discarding deleted columns from $\boldsymbol{D}^{(1)}$
   - Grouping replicas in $\boldsymbol{D}^{(2)}$ by marker addition



3. Check if X^n and $\tilde{Y}$ are jointly typical.
4. Match $\tilde{Y}$ with X^n, if X^n is the only row jointly typical with $\tilde{Y}$.

VII. Converse

❖ A genie-aided proof
- The availability of the repetition pattern S^n is assumed

❖ Provides insight into privacy-preserving anonymized data sharing/publication

VIII. Conclusion

❖ A unified framework of database matching in the presence of noise and synchronization errors

❖ Repetition pattern can be extracted using seeds.
- A seed size linear with the column size is sufficient.
- Seed size is logarithmic with the number of users!

❖ Replicas increase matching capacity if detected correctly!
- Replicas behave as "repetition code of varying length"

References

1. S. Bakirtas and E. Erkip, "Database Matching Under Column Deletions", ISIT 2021
2. S. Bakirtas and E. Erkip, "Seeded Database Matching Under Noisy Column Repetitions", to appear in ITW 2022
3. S. Bakirtas and E. Erkip, "Matching of Markov Databases Under Random Column Repetitions", to appear in Asilomar 2022