Matching of Markov Databases Under Random Column Repetitions

1354

NYU TANDON SCHOOL OF ENGINEERING

NYU WIRELESS

Serhat Bakirtas, Elza Erkip
{serhat.bakirtas,elza}@nyu.edu

2022 Asilomar Conference on Signals, Systems, and Computers

Introduction

- Data collection is booming.
- Personal microdata are published after anonymization.
- Anonymized data are not truly private.
- Correlated public data can be exploited for de-anonymization!
  - Database Matching

Applications of Database Matching

- Data & network privacy
- Computer vision
- DNA sequencing
- Single-cell biological data alignment

System Model

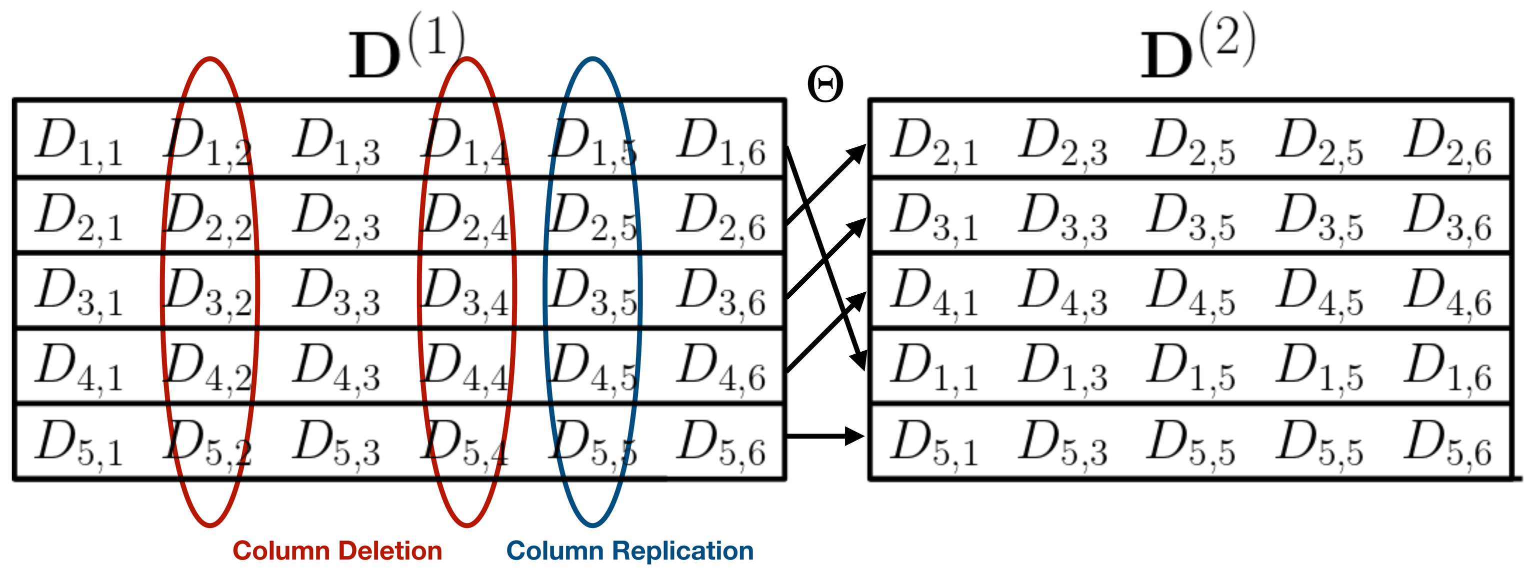
- Unlabeled Database: $\mathbf{D}^{(1)} \in \mathfrak{X}^{m_n \times n}$ with
  - I.I.D. rows following a first-order stationary Markov process capturing correlation among attributes.
  - Probability transition matrix $\mathbf{P}$ over $\mathfrak{X}$

$$\mathbf{P} = \gamma \mathbf{I} + (1-\gamma)\mathbf{U}$$
$$U_{i,j} = u_j > 0, \ \forall (i,j) \in \mathfrak{X}^2$$

  - π: stationary distribution of $\mathbf{P}$.
- Labeling Function: Uniform permutation Θ_n of $[m_n]$.
- Synchronization Errors: Random column repetition pattern $S^n \overset{\text{i.i.d.}}{\sim} p_S$, $\delta \triangleq p_S(0)$.
- Labeled Database: Pair $(\mathbf{D}^{(2)}, \Theta_n)$ with

$$D^{(2)}_{i,j} = \begin{cases} E, & \text{if } S_j = 0 \\ D^{(1)}_{\Theta_n^{-1}(i),j} \otimes \mathbb{1}^{S_j} & \text{if } S_j \geq 1 \end{cases}$$

- Database Growth Rate: $R = \lim_{n\to\infty} \frac{\log_2 m_n}{n}$.
- Matching: Estimation of Θ_n.



Objectives

- What are the **sufficient** and the **necessary** conditions on the database growth rate for successful matching?
- Can we infer the repetition pattern S^n from $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$? If yes, how?

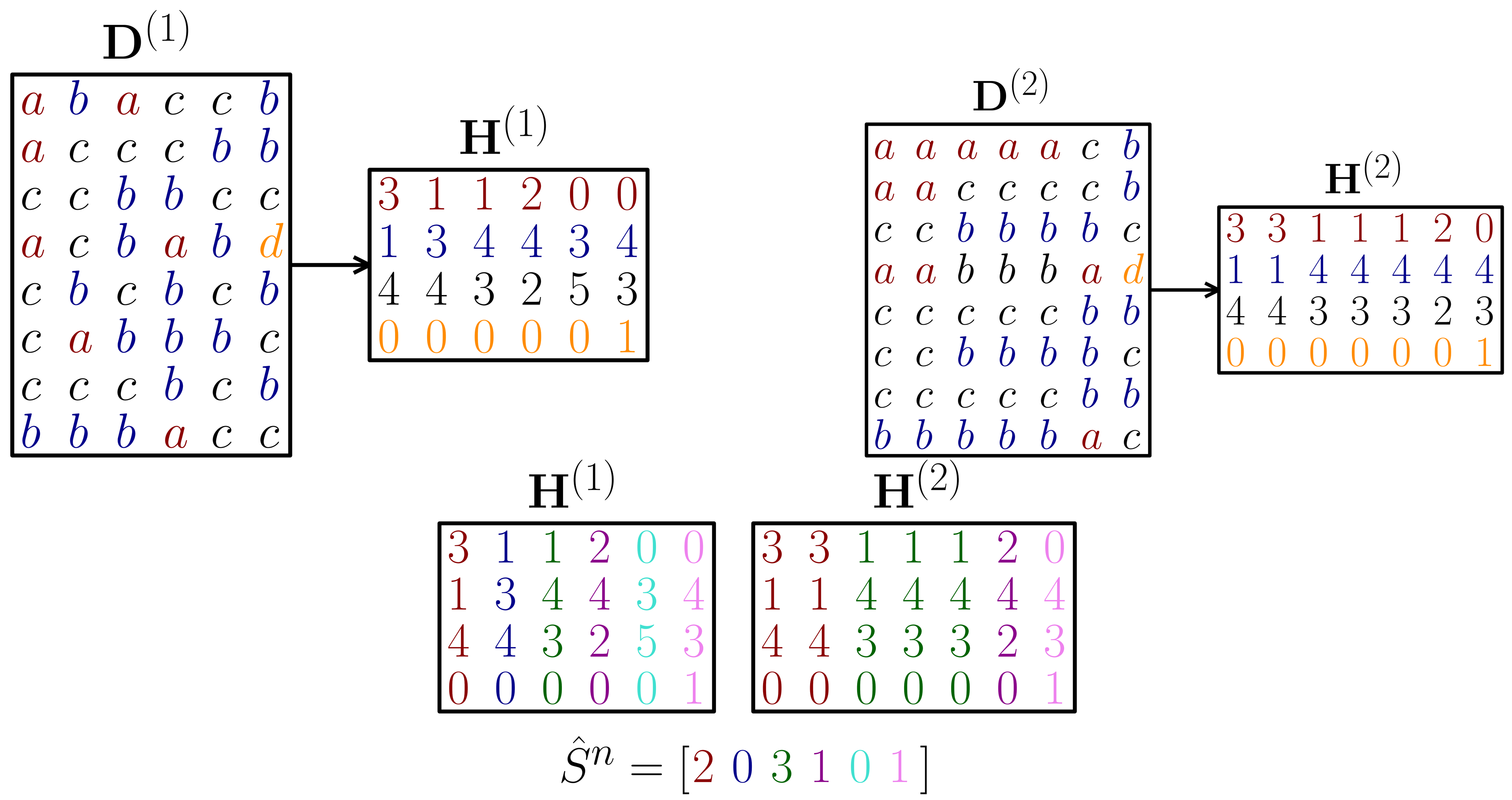
Main Result

Databases with growth rate R can be successfully matched if $R < C$ where

$$C \triangleq \frac{(1-\delta)(1-\gamma)}{(1-\gamma\delta)}[H(\pi) + \sum_{i\in\mathfrak{X}} u_i^2 \log u_i]$$
$$-(1-\delta)^2 \sum_{r=0}^{\infty} \delta^r \sum_{i\in\mathfrak{X}} u_i(\gamma^{r+1} + (1-\gamma^{r+1})u_i)\log(\gamma^{r+1} + (1-\gamma^{r+1})u_i)$$

Furthermore, a necessary condition for the existence of a successful matching scheme is $R \leq C$.

Achievability-I: Histogram-Based Repetition Detection

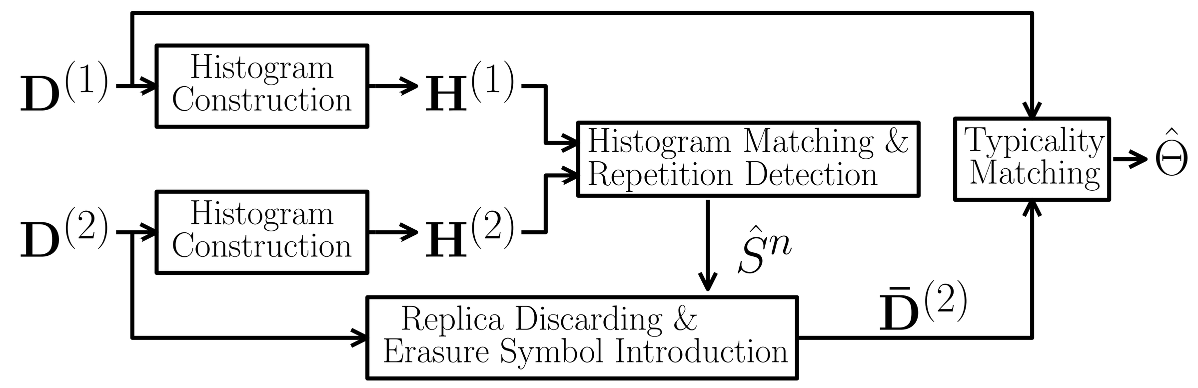


Lemma: Asymptotic Uniqueness of The Histograms

As long as $m_n = \omega(n^4)$

- column histograms are asymptotically unique
- histogram-based repetition detection is successful.

Achievability-II: Matching Scheme



Converse

- A genie aided proof, assuming the repetition pattern S^n.
- Provides insight into privacy-preserving anonymized data sharing/publication

Conclusion

- A wide range of applications of database matching
- Existence of an underlying structure helps.
- Column histograms of the databases are asymptotically unique.
- Histograms help us infer the repetition pattern.
- A tight bound on the achievable database growth rates.