Matching of Markov Databases Under Random Column Repetitions

Serhat Bakirtas, Elza Erkip

New York University

2022 Asilomar Conference on Signals, Systems, and Computers

Motivation

- Age of data collection.

Motivation

- Age of data collection.
- Potentially-sensitive data are made available for commercial and research purposes.

Motivation

- Age of data collection.
- Potentially-sensitive data are made available for commercial and research purposes.
  - User identities are removed: *Anonymization*.

Motivation

- Age of data collection.
- Potentially-sensitive data are made available for commercial and research purposes.
  - User identities are removed: *Anonymization*.
- Are anonymized data truly private?

Motivation

- Age of data collection.
- Potentially-sensitive data are made available for commercial and research purposes.
  - User identities are removed: *Anonymization*.
- Are anonymized data truly private?
- NO!

Motivation

- Age of data collection.
- Potentially-sensitive data are made available for commercial and research purposes.
  - User identities are removed: *Anonymization*.
- Are anonymized data truly private?
- NO!
  - Correlated public data $\rightarrow$ De-anonymization!

We Found Joe Biden's Secret Venmo. Here's Why That's A Privacy Nightmare For Everyone.

The peer-to-peer payments app leaves everyone from ordinary people to the most powerful person in the world exposed.


Ryan Mac
BuzzFeed News Reporter


Katie Notopoulos
BuzzFeed News Reporter


Ryan Brooks
BuzzFeed News Reporter

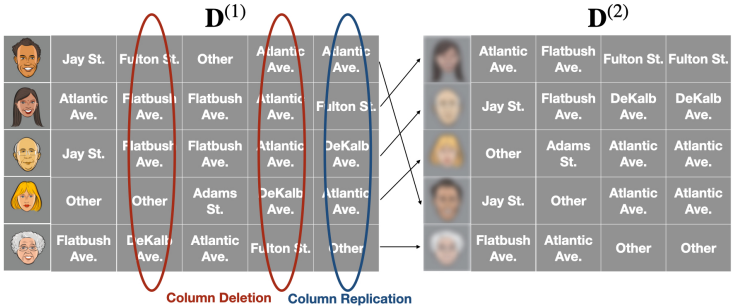

Logan McDonald
BuzzFeed Staff

Motivation: Our Work

- Anonymized databases containing *micro-information* shared and published routinely.
- Examples: Movie preferences, financial transactions data, location data, health records.

Motivation: Our Work

- Anonymized databases containing *micro-information* shared and published routinely.
- Examples: Movie preferences, financial transactions data, location data, health records.
- This work: Time-indexed data, *e.g.*, financial and location data
- Synchronization errors in time-indexed data: column repetitions

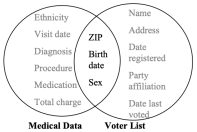


Practical Database Matching Attacks

- [Narayanan and Shmatikov, 2008]
  De-anonymization of Netflix Prize Database using IMDB data.

- [Sweeney, 2002]
  De-anonymization of medical databases using voter registration data.

- [Naini et al., 2012]
  User identification from geolocation data.



(a) Unlabeled histograms (Day 1)

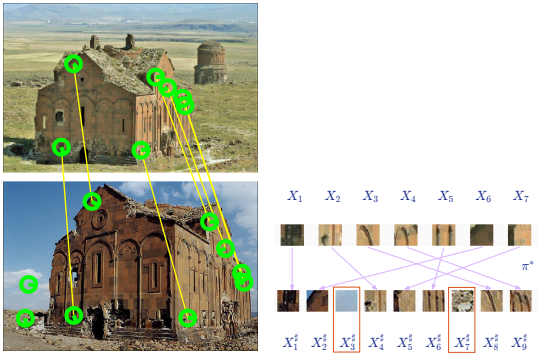
User	Location		
	Dorm.	Rest.	Lib.
?	75%	15%	10%
?	31%	30%	39%
?	15%	15%	70%
?	15%	65%	20%

(b) Labeled histograms (Day 2)

User	Location		
	Dorm.	Rest.	Lib.
John	33%	33%	34%
Jill	70%	20%	10%
Mary	15%	60%	25%
Mike	15%	20%	65%

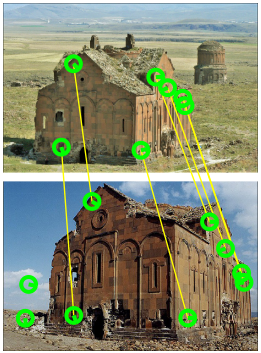
Database Matching: Other Applications

- Computer vision [Galstyan et al., 2021]



Database Matching: Other Applications

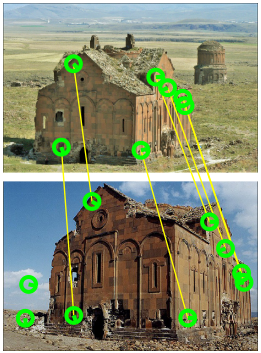
- Computer vision [Galstyan et al., 2021]

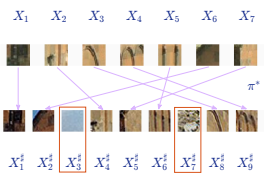


- Biological applications
  - DNA Sequencing [Blazewicz et al., 2002]

Database Matching: Other Applications

- Computer vision [Galstyan et al., 2021]

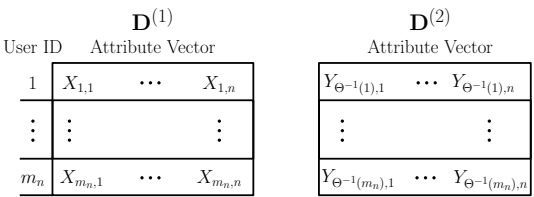




- Biological applications
  - DNA Sequencing [Blazewicz et al., 2002]
  - Single-cell data alignment [Chen et al., 2022]

Previous Works: Information-Theoretical Limits

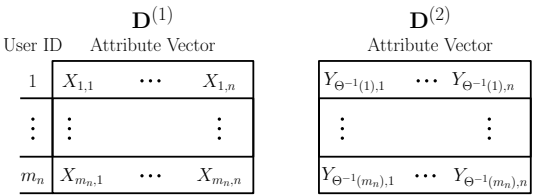
[Shirani, Garg, and Erkip, ISIT 2019]



- Databases as $m_n \times n$ random matrices: equal no. of labeled attributes (columns)
  - Matching rows $\sim f_{X^{(1),n},X^{(2),n}}$

Previous Works: Information-Theoretical Limits

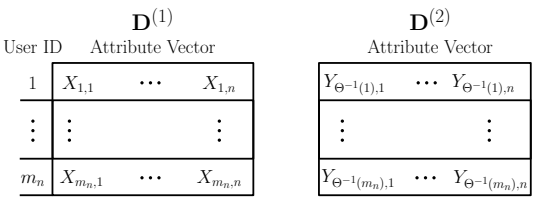
[Shirani, Garg, and Erkip, ISIT 2019]



- Databases as $m_n \times n$ random matrices: equal no. of labeled attributes (columns)
  - Matching rows $\sim f_{X^{(1),n},X^{(2),n}}$
- Database growth rate: $R = \lim_{n\to\infty} \frac{1}{n} \log m_n$

Previous Works: Information-Theoretical Limits

[Shirani, Garg, and Erkip, ISIT 2019]



- Databases as $m_n \times n$ random matrices: equal no. of labeled attributes (columns)
  - Matching rows $\sim f_{X^{(1),n},X^{(2),n}}$
- Database growth rate: $R = \lim_{n\to\infty} \frac{1}{n} \log m_n$
- Successful matching: $P_e \to 0$ as $n \to \infty$
- Database matching $\Leftrightarrow$ Channel decoding

Previous Works: Information-Theoretical Limits

- Objective: Given $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$, find $\hat{\Theta}$ s.t.:

$$\Pr(\Theta(I) = \hat{\Theta}(I)) \to 1 \text{ as } n \to \infty,$$

  where $I \sim U(1, m_n)$.

- Almost all entries must be matched correctly.

Previous Works: Information-Theoretical Limits

- Objective: Given $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$, find $\hat{\Theta}$ s.t.:

$$\Pr(\Theta(I) = \hat{\Theta}(I)) \to 1 \text{ as } n \to \infty,$$

  where $I \sim U(1, m_n)$.

- Almost all entries must be matched correctly.
  - In [Cullina et al., 2018], [Dai et al., 2019]: All entries must be matched correctly.

Previous Works: Information-Theoretical Limits

- Objective: Given $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$, find $\hat{\Theta}$ s.t.:

$$\Pr(\Theta(I) = \hat{\Theta}(I)) \to 1 \text{ as } n \to \infty,$$

  where $I \sim U(1, m_n)$.

- Almost all entries must be matched correctly.
  - In [Cullina et al., 2018], [Dai et al., 2019]: All entries must be matched correctly.
- This allows us to
  - use information-theoretic tools,
  - work with arbitrary distributions.

Previous Works: Information-Theoretical Limits

[Bakirtas and Erkip, ISIT 2021]

- Database Matching Under Column Deletions.
  - Different number of attributes (columns).
  - Attributes are unlabeled.

Previous Works: Information-Theoretical Limits

[Bakirtas and Erkip, ISIT 2021]

- Database Matching Under Column Deletions.
  - Different number of attributes (columns).
  - Attributes are unlabeled.

- Sufficient conditions on R for database matching
  - Side information on the deletion locations.

Previous Works: Information-Theoretical Limits

[Bakirtas and Erkip, ISIT 2021]

- Database Matching Under Column Deletions.
  - Different number of attributes (columns).
  - Attributes are unlabeled.
- Sufficient conditions on R for database matching
  - Side information on the deletion locations.
- Extracting this side information from a batch of correctly-matched rows (seeds).

Previous Works: Information-Theoretical Limits

[Bakirtas and Erkip, ITW 2022]

- Seeded Database Matching Under Noisy Random Column Repetitions.
  - Different number of attributes (columns).
  - Attributes are unlabeled.
  - Noisy entries.
  - Seeds available.

Previous Works: Information-Theoretical Limits

[Bakirtas and Erkip, ITW 2022]

- Seeded Database Matching Under Noisy Random Column Repetitions.
  - Different number of attributes (columns).
  - Attributes are unlabeled.
  - Noisy entries.
  - Seeds available.

- Noisy repetition detection algorithms exploiting seeds.

Previous Works: Information-Theoretical Limits

[Bakirtas and Erkip, ITW 2022]

- Seeded Database Matching Under Noisy Random Column Repetitions.
  - Different number of attributes (columns).
  - Attributes are unlabeled.
  - Noisy entries.
  - Seeds available.
- Noisy repetition detection algorithms exploiting seeds.
- Sufficient conditions on R for database matching given $\Omega(n)$ seeds.

Previous Works: Information-Theoretical Limits

[Bakirtas and Erkip, ITW 2022]

- Seeded Database Matching Under Noisy Random Column Repetitions.
  - Different number of attributes (columns).
  - Attributes are unlabeled.
  - Noisy entries.
  - Seeds available.
- Noisy repetition detection algorithms exploiting seeds.
- Sufficient conditions on R for database matching given $\Omega(n)$ seeds.
- Necessary conditions on R for database matching.

Matching of Markov Databases Under Random Column Repetitions

We assume

1. The attributes are not labeled.
2. Databases do not have the same number of attributes.
   - Random column repetitions.

Matching of Markov Databases Under Random Column Repetitions

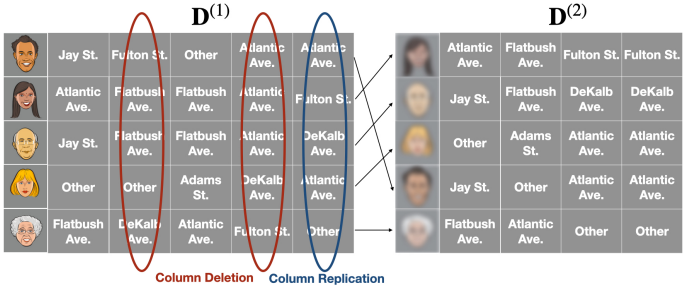
We assume

1. The attributes are not labeled.
2. Databases do not have the same number of attributes.
   - Random column repetitions.
3. The indices of the repeated columns are not known.

Matching of Markov Databases Under Random Column Repetitions

We assume

1. The attributes are not labeled.
2. Databases do not have the same number of attributes.
   - Random column repetitions.
3. The indices of the repeated columns are not known.
4. Repetition pattern is constant across rows.
   - Generalization to non-constant repetition pattern is possible.



System Model

- $\mathbf{D}^{(1)}$: $m_n \times n$ random matrix.
  - Rows are *i.i.d.*

System Model

- $\mathbf{D}^{(1)}$: $m_n \times n$ random matrix.
  - Rows are *i.i.d.*
  - Each row follows a first-order stationary Markov process with probability transition matrix $\mathbf{P}$ over $\mathfrak{X}$ such that

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

System Model

- $\mathbf{D}^{(1)}$: $m_n \times n$ random matrix.
  - Rows are *i.i.d.*
  - Each row follows a first-order stationary Markov process with probability transition matrix $\mathbf{P}$ over $\mathfrak{X}$ such that

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

  - Markov: Correlation among attributes.
  - $D_{i,1} \overset{\text{i.i.d.}}{\sim} \pi = [u_1, \ldots, u_{|\mathfrak{X}|}],\ i = 1, \ldots, m_n$.

System Model

- $\mathbf{D}^{(1)}$: $m_n \times n$ random matrix.
  - Rows are *i.i.d.*
  - Each row follows a first-order stationary Markov process with probability transition matrix $\mathbf{P}$ over $\mathfrak{X}$ such that

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

  - Markov: Correlation among attributes.
  - $D_{i,1} \overset{\text{i.i.d.}}{\sim} \pi = [u_1, \ldots, u_{|\mathfrak{X}|}],\ i = 1, \ldots, m_n.$
  - π is the stationary distribution associated with $\mathbf{P}$.

System Model

- $\mathbf{D}^{(1)}$: $m_n \times n$ random matrix.
  - Rows are *i.i.d.*
  - Each row follows a first-order stationary Markov process with probability transition matrix $\mathbf{P}$ over $\mathfrak{X}$ such that

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

  - Markov: Correlation among attributes.
  - $D_{i,1} \overset{\text{i.i.d.}}{\sim} \pi = [u_1, \ldots, u_{|\mathfrak{X}|}],\ i = 1, \ldots, m_n$.
  - π is the stationary distribution associated with $\mathbf{P}$.
- Database Growth Rate: $R = \lim_{n\to\infty} \frac{1}{n} \log m_n$
  - Assumption: $R > 0$.

System Model

- $\mathbf{D}^{(1)}$: $m_n \times n$ random matrix.
  - Rows are *i.i.d.*
  - Each row follows a first-order stationary Markov process with probability transition matrix $\mathbf{P}$ over $\mathfrak{X}$ such that

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

  - Markov: Correlation among attributes.
  - $D_{i,1} \overset{\text{i.i.d.}}{\sim} \pi = [u_1, \ldots, u_{|\mathfrak{X}|}]$, $i = 1, \ldots, m_n$.
  - π is the stationary distribution associated with $\mathbf{P}$.
- Database Growth Rate: $R = \lim_{n\to\infty} \frac{1}{n} \log m_n$
  - Assumption: $R > 0$.
- Θ: Uniform permutation of $[m_n]$.

System Model: Continued

- Column repetition pattern: random vector $S^n = \{S_1, S_2, ..., S_n\}$ with $S_j \overset{i.i.d.}{\sim} p_S$.
  - $supp(p_S) = \{0, \ldots, s_{\max}\}$
- $\mathbf{D}^{(2)}$: Obtained from $\mathbf{D}^{(1)}$ by
  1. Row shuffling by Θ.

System Model: Continued

- Column repetition pattern: random vector $S^n = \{S_1, S_2, ..., S_n\}$ with $S_j \overset{i.i.d.}{\sim} p_S$.
  - $supp(p_S) = \{0, \ldots, s_{\max}\}$
- $\mathbf{D}^{(2)}$: Obtained from $\mathbf{D}^{(1)}$ by
  1. Row shuffling by Θ.
  2. Column repetition by S^n.
     - Replicate the j^{th} column S_j times if $S_j > 0$.
     - Delete the j^{th} column if $S_j = 0$.

System Model: Continued

- Column repetition pattern: random vector $S^n = \{S_1, S_2, ..., S_n\}$ with $S_j \overset{i.i.d.}{\sim} p_S$.
  - $supp(p_S) = \{0, \ldots, s_{\max}\}$
- $\mathbf{D}^{(2)}$: Obtained from $\mathbf{D}^{(1)}$ by
  1. Row shuffling by Θ.
  2. Column repetition by S^n.
     - Replicate the j^{th} column S_j times if $S_j > 0$.
     - Delete the j^{th} column if $S_j = 0$.
- No noise on the entries.

System Model: Continued

- Achievable Database Growth Rate: Given $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$ with growth rate R, $\exists \hat{\Theta}$ s.t.:

$$\Pr(\Theta(I) = \hat{\Theta}(I)) \to 1 \text{ as } n \to \infty,$$

where $I \sim U(1, m_n)$.

- Matching Capacity:

$$C \triangleq \sup\{R\colon\ R \text{ is achievable.}\}$$

System Model: Continued

- Achievable Database Growth Rate: Given $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$ with growth rate R, $\exists \hat{\Theta}$ s.t.:

$$\Pr(\Theta(I) = \hat{\Theta}(I)) \to 1 \text{ as } n \to \infty,$$

where $I \sim U(1, m_n)$.

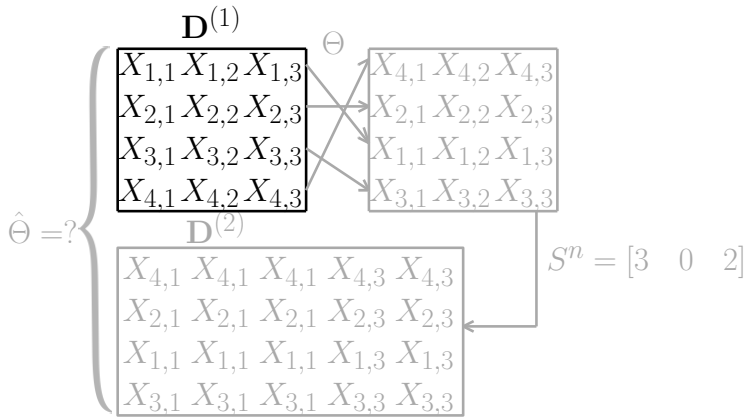
- Matching Capacity:

$$C \triangleq \sup\{R\colon\ R \text{ is achievable.}\}$$

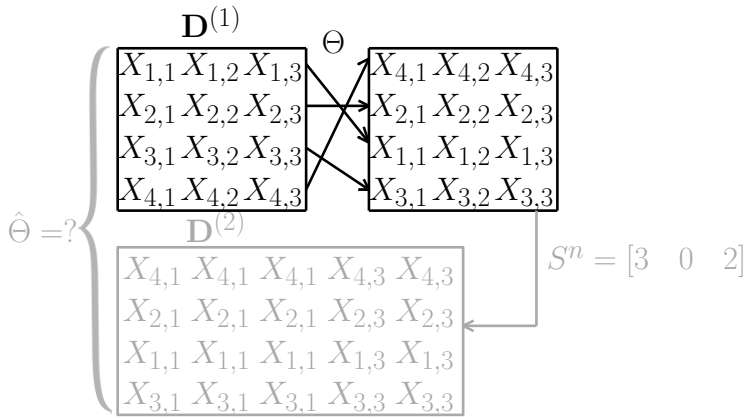
- Goal: Given $\mathbf{P}$, characterize matching capacity C.

System Model: Example

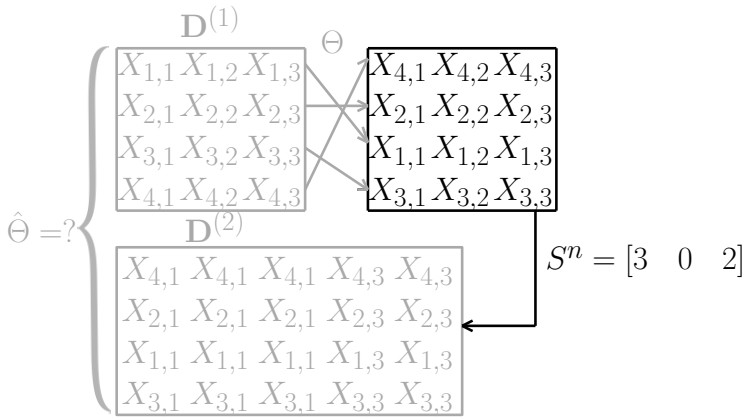
- $m_n = 4, n = 3$, S^n: repetition pattern.



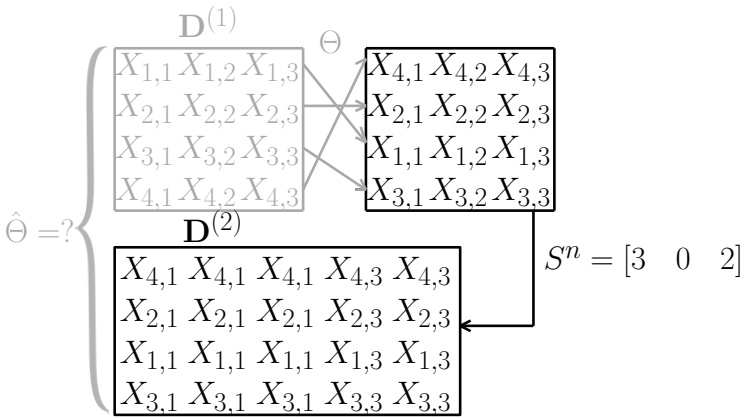
System Model: Example



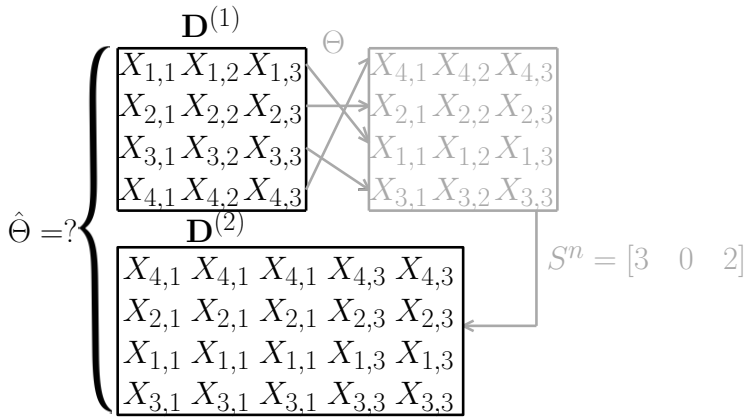
System Model: Example



System Model: Example



System Model: Example



This Talk: Objectives

1. What is the matching capacity?

This Talk: Objectives

1. What is the matching capacity?
2. Can we devise matching schemes which achieve this matching capacity?

This Talk: Objectives

1. What is the matching capacity?
2. Can we devise matching schemes which achieve this matching capacity?
3. Can we infer the repetition pattern S^n from $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$ without any seeds?

This Talk: Objectives

1. What is the matching capacity?
2. Can we devise matching schemes which achieve this matching capacity?
3. Can we infer the repetition pattern S^n from $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$ without any seeds?
4. If yes, how?

Proposed Matching Scheme

- Exploit the identical repetition pattern across rows.

Proposed Matching Scheme

- Exploit the identical repetition pattern across rows.
  1. Find a permutation-invariant unique feature of the columns.

Proposed Matching Scheme

- Exploit the identical repetition pattern across rows.
  1. Find a permutation-invariant unique feature of the columns.
  2. By matching these features, infer S^n.

Proposed Matching Scheme

- Exploit the identical repetition pattern across rows.
  1. Find a permutation-invariant unique feature of the columns.
  2. By matching these features, infer S^n.
  3. Replace the deleted columns with erasure symbol $*$ in $\mathbf{D}^{(2)}$.
  4. Discard the replicated columns from $\mathbf{D}^{(2)}$.

Proposed Matching Scheme

- Exploit the identical repetition pattern across rows.
  1. Find a permutation-invariant unique feature of the columns.
  2. By matching these features, infer S^n.
  3. Replace the deleted columns with erasure symbol $*$ in $\mathbf{D}^{(2)}$.
  4. Discard the replicated columns from $\mathbf{D}^{(2)}$.
  5. Perform typicality-based rowwise matching, with respect to Erasure Channel ($p_S(0)$).

Proposed Matching Scheme

- Exploit the identical repetition pattern across rows.
  1. Find a permutation-invariant unique feature of the columns.
  2. By matching these features, infer S^n.
  3. Replace the deleted columns with erasure symbol $*$ in $\mathbf{D}^{(2)}$.
  4. Discard the replicated columns from $\mathbf{D}^{(2)}$.
  5. Perform typicality-based rowwise matching, with respect to Erasure Channel ($p_S(0)$).

- We will use column histograms as the permutation-invariant feature.

Proposed Matching Scheme

- Exploit the identical repetition pattern across rows.
  1. Find a permutation-invariant unique feature of the columns.
  2. By matching these features, infer S^n.
  3. Replace the deleted columns with erasure symbol $*$ in $\mathbf{D}^{(2)}$.
  4. Discard the replicated columns from $\mathbf{D}^{(2)}$.
  5. Perform typicality-based rowwise matching, with respect to Erasure Channel ($p_S(0)$).

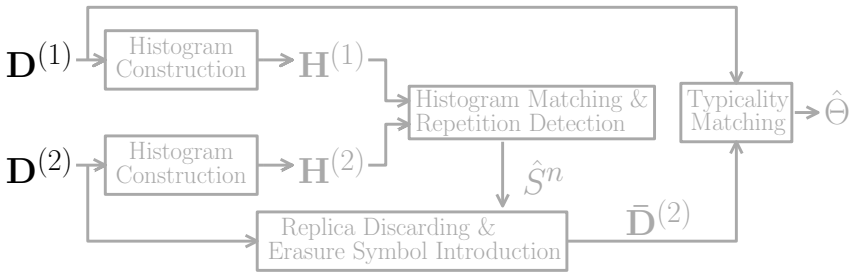
- We will use column histograms as the permutation-invariant feature.
- As long as the column histograms are unique, we can
  - Label each attribute correctly.

Proposed Matching Scheme

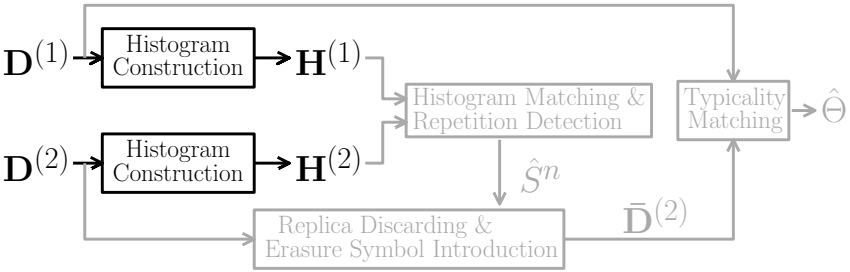
- Exploit the identical repetition pattern across rows.
  1. Find a permutation-invariant unique feature of the columns.
  2. By matching these features, infer S^n.
  3. Replace the deleted columns with erasure symbol $*$ in $\mathbf{D}^{(2)}$.
  4. Discard the replicated columns from $\mathbf{D}^{(2)}$.
  5. Perform typicality-based rowwise matching, with respect to Erasure Channel ($p_S(0)$).

- We will use column histograms as the permutation-invariant feature.
- As long as the column histograms are unique, we can
  - Label each attribute correctly.
  - Infer the repetition pattern.

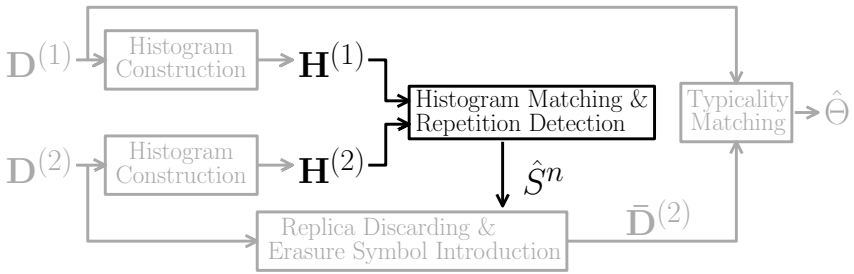
Matching Scheme



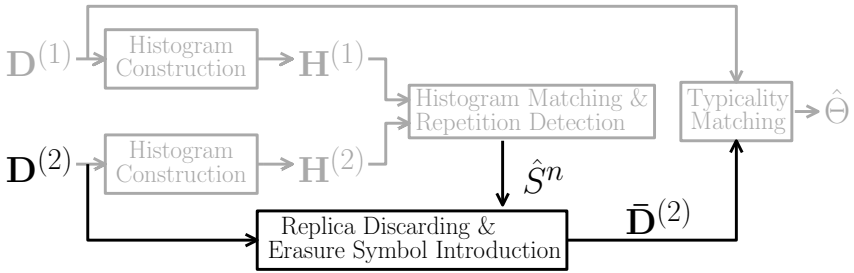
Matching Scheme



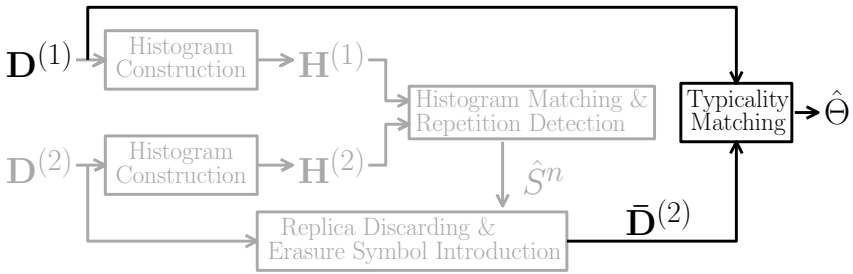
Matching Scheme



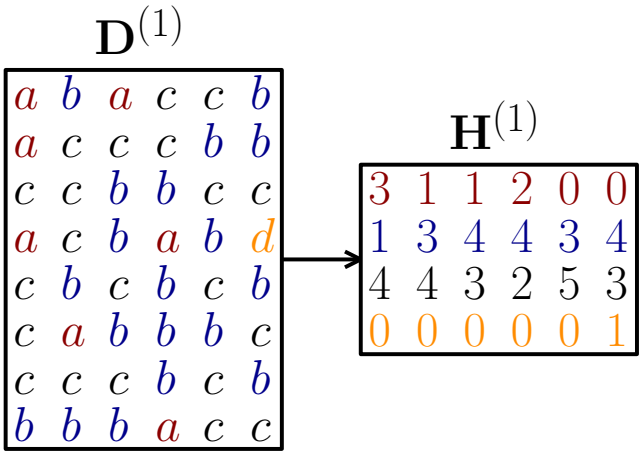
Matching Scheme



Matching Scheme



Histogram-Based Repetition Detection: Example



Histogram-Based Repetition Detection: Example

$\mathbf{D}^{(2)}$

a	a	a	a	a	c	b
a	a	c	c	c	c	b
c	c	b	b	b	b	c
a	a	b	b	b	a	d
c	c	c	c	c	b	b
c	c	b	b	b	b	c
c	c	c	c	c	b	b
b	b	b	b	b	a	c

$\rightarrow$

$\mathbf{H}^{(2)}$

3	3	1	1	1	2	0
1	1	4	4	4	4	4
4	4	3	3	3	2	3
0	0	0	0	0	0	1

Histogram-Based Repetition Detection: Example

$$\mathbf{H}^{(1)} = \begin{bmatrix} 3 & 1 & 1 & 2 & 0 & 0 \\ 1 & 3 & 4 & 4 & 3 & 4 \\ 4 & 4 & 3 & 2 & 5 & 3 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{H}^{(2)} = \begin{bmatrix} 3 & 3 & 1 & 1 & 1 & 2 & 0 \\ 1 & 1 & 4 & 4 & 4 & 4 & 4 \\ 4 & 4 & 3 & 3 & 3 & 2 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\hat{S}^n = [2\ 0\ 3\ 1\ 0\ 1]$$

Asymptotic Uniqueness of The Histograms

Lemma

Let H_i denote the i$^{\text{th}}$ column of the histogram matrix $\mathbf{H}^{(1)}$. Then, $\Pr(\exists i, j \in [n],\ i \neq j, H_i = H_j) \to 0$ as $n \to \infty$ if $m_n = \omega\left(n^4\right)$.

Asymptotic Uniqueness of The Histograms

Lemma

Let H_i denote the i$^{\text{th}}$ column of the histogram matrix $\mathbf{H}^{(1)}$. Then, $\Pr(\exists i, j \in [n],\ i \neq j, H_i = H_j) \to 0$ as $n \to \infty$ if $m_n = \omega\left(n^4\right)$.

- For $R > 0$, $m_n = \omega(n^p)\ \forall p \in \mathbb{N}$.

$\Rightarrow$ Asymptotically, columns of $\mathbf{H}^{(1)}$ are unique.

Asymptotic Uniqueness of The Histograms

Lemma

Let H_i denote the i$^{\text{th}}$ column of the histogram matrix $\mathbf{H}^{(1)}$. Then, $\Pr(\exists i, j \in [n],\ i \neq j, H_i = H_j) \to 0$ as $n \to \infty$ if $m_n = \omega\left(n^4\right)$.

- For $R > 0$, $m_n = \omega(n^p)\ \forall p \in \mathbb{N}$.

$\Rightarrow$ Asymptotically, columns of $\mathbf{H}^{(1)}$ are unique.

- Since there is no noise, they can be matched with the columns of $\mathbf{H}^{(2)}$.

Lemma: Sketch of Proof

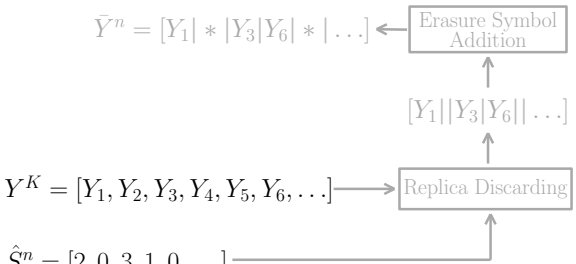
- "Collapse" the databases to binary ones.
- Union bound

$$\Pr(\exists i,j \in [n],\ i \neq j, H_i = H_j) \leq \sum_{(i,j)\in[n]^2:i<j} \Pr(\tilde{H}_i^{(1)} = \tilde{H}_j^{(1)})$$
$$\leq n^2 \max_{(i,j)\in[n]^2:i<j} \Pr(\tilde{H}_i^{(1)} = \tilde{H}_j^{(1)})$$

- $\Pr(\tilde{H}_i^{(1)} = \tilde{H}_j^{(1)}) = \sum_{r=0}^{m_n} \binom{m}{r}(1-u_1)^r u_1^{m_n - r} \Pr(\tilde{H}_j^{(1)} = r | \tilde{H}_1^{(1)} = r)$
- $\Pr(\tilde{H}_j^{(1)} = r | \tilde{H}_1^{(1)} = r)$ is $\Pr(A + B = r)$, where A and B are Binomials.
- Apply Stirling's approximation and the method of types
- Separate into two cases based on the type of Binomials
  - If not typical, summands decay exponentially
  - If typical, apply Pinsker's inequality
- Choose parameters carefully.

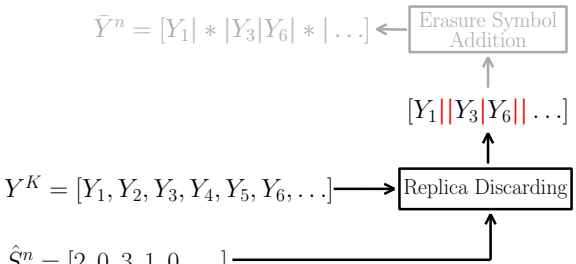
Replica Discarding & Erasure Symbol Addition: Example

Y^K: a row of $\mathbf{D}^{(2)}$



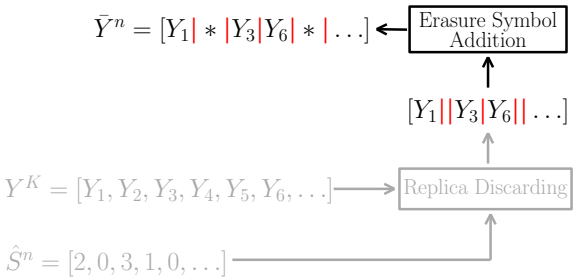
Replica Discarding & Erasure Symbol Addition: Example

Y^K: a row of $\mathbf{D}^{(2)}$



Replica Discarding & Erasure Symbol Addition: Example

Y^K: a row of $\mathbf{D}^{(2)}$



Finally, check the joint typicality of $(X^n, \bar{Y}^n)$.

Main Result: Matching Capacity

Theorem

Given a probability transition matrix $\mathbf{P}$ and a repetition probability distribution p_S, the matching capacity is

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

where $\delta \triangleq p_S(0)$.

Main Result: Continued

- Any deletion/replica can be converted to erasure.

Main Result: Continued

- Any deletion/replica can be converted to erasure.
- Deleted/erased columns offer temporal information.

Main Result: Continued

- Any deletion/replica can be converted to erasure.
- Deleted/erased columns offer temporal information.
- Replicated columns do not offer additional information.

Main Result: Continued

- Any deletion/replica can be converted to erasure.
- Deleted/erased columns offer temporal information.
- Replicated columns do not offer additional information.
- We have a complete characterization of the matching capacity.

Conclusion

- Database Matching $\Leftrightarrow$ Channel Decoding

Conclusion

- Database Matching $\Leftrightarrow$ Channel Decoding
- Existence of an underlying structure helps.

Conclusion

- Database Matching $\Leftrightarrow$ Channel Decoding
- Existence of an underlying structure helps.
- Column histograms of the databases are asymptotically unique.

Conclusion

- Database Matching $\Leftrightarrow$ Channel Decoding
- Existence of an underlying structure helps.
- Column histograms of the databases are asymptotically unique.
- Histograms help us infer the repetition pattern.

Conclusion

- Database Matching $\Leftrightarrow$ Channel Decoding
- 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.

Conclusion

- Database Matching $\Leftrightarrow$ Channel Decoding
- 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.
- Converse result $\Rightarrow$ Insight into privacy-preserving publication of anonymized time-indexed microdata.

Conclusion

- Database Matching $\Leftrightarrow$ Channel Decoding
- 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.
- Converse result $\Rightarrow$ Insight into privacy-preserving publication of anonymized time-indexed microdata.
- Ongoing Work: Database matching when the repetition pattern is not constant across rows.

Thank you! Q&A?

Matching of Markov Databases Under Random Column Repetitions

Serhat Bakirtas, Elza Erkip

serhat.bakirtas@nyu.edu