Database Matching Under Adversarial Column Deletions

Serhat Bakirtas, Elza Erkip

New York University

2023 IEEE Information Theory Workshop

Saint-Malo, France

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.

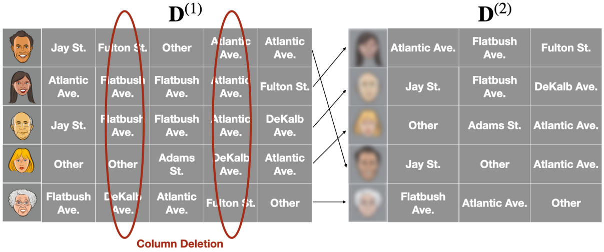
Ryan Mac BuzzFeed News Reporter
Katie Notopoulos BuzzFeed News Reporter
Ryan Brooks BuzzFeed News Reporter
Logan McDonald BuzzFeed Staff

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: De-anonymization of time-indexed data, *e.g.*, financial and location data



Motivation: Loss of Synchronization in Time-Indexed Data

Loss of synchronization in time-indexed data, due to

1. Sampling errors

Motivation: Loss of Synchronization in Time-Indexed Data

Loss of synchronization in time-indexed data, due to

1. Sampling errors
   - Random column deletions & replications

Motivation: Loss of Synchronization in Time-Indexed Data

Loss of synchronization in time-indexed data, due to

1. Sampling errors
   - Random column deletions & replications
2. A privacy-preserving mechanism

Motivation: Loss of Synchronization in Time-Indexed Data

Loss of synchronization in time-indexed data, due to

1. Sampling errors
   - Random column deletions & replications
2. A privacy-preserving mechanism
   - Intentional/Adversarial! column deletions

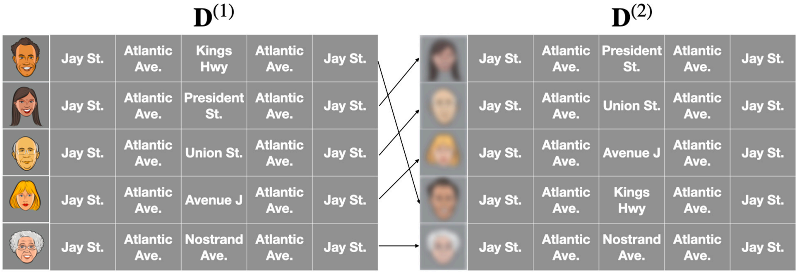
Motivation: Loss of Synchronization in Time-Indexed Data

Loss of synchronization in time-indexed data, due to

1. Sampling errors
   - Random column deletions & replications
2. A privacy-preserving mechanism
   - Intentional/Adversarial! column deletions
   - A deletion budget: Privacy - Utility trade-off

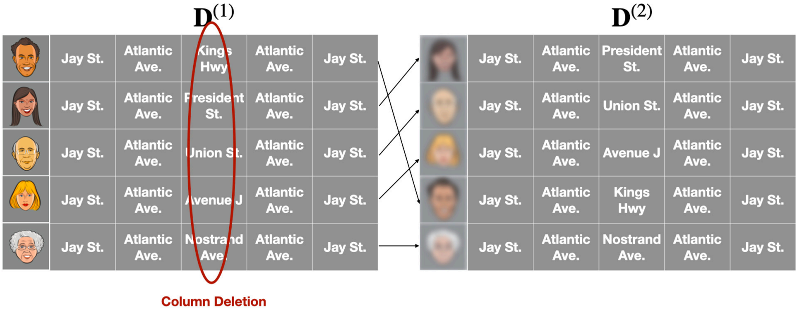
Motivation: Adversarial Column Deletions

- Some time-instances may offer more information than others.
  - *e.g.* Night-time locations reveal more private information.



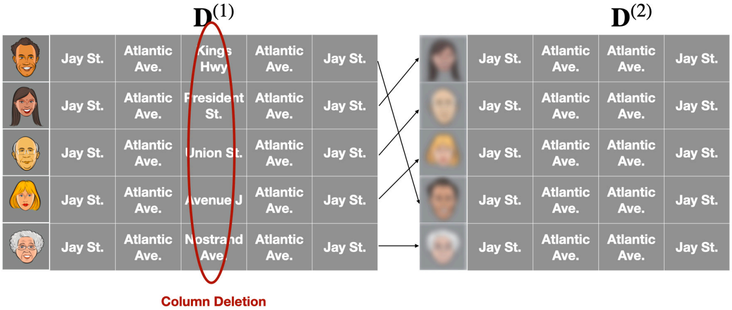
Motivation: Adversarial Column Deletions

- Some time-instances may offer more information than others.
  - *e.g.* Night-time locations reveal more private information.



Motivation: Adversarial Column Deletions

- Some time-instances may offer more information than others.
  - *e.g.* Night-time locations reveal more private information.

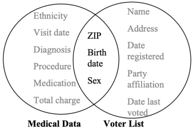


Practical Database De-Anonymization Attacks

- [Narayanan and Shmatikov, 2008]
  De-anonymization of Netflix Prize Dataset 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)

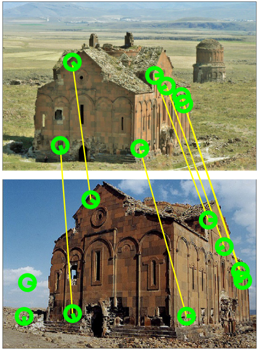
User	Location		
	Dorm.	Rest.	Lib.
?	75%	15%	10%
?	31%	30%	39%
?	15%	15%	70%
?	15%	65%	20%

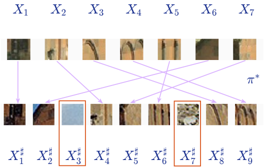
(b) Labeled histograms (Day 2)

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

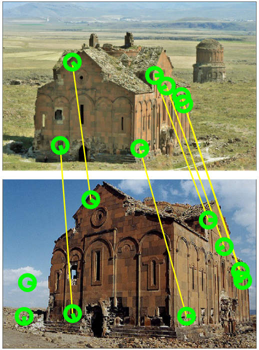
- Computer vision [Galstyan et al., 2021]

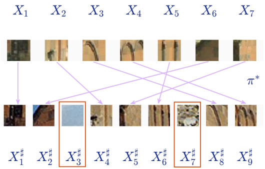




Database Matching: Other Applications

- Computer vision [Galstyan et al., 2021]

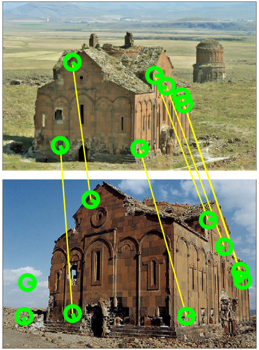


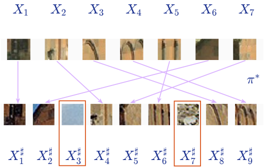


- Biological applications
  - DNA Sequencing [Blazewicz et al., 2002]

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 '19]

$\mathbf{D}^{(1)}$

User ID	Attribute Vector		
1	$X_{1,1}$	$\cdots$	$X_{1,n}$
$\vdots$	$\vdots$		$\vdots$
m_n	$X_{m_n,1}$	$\cdots$	$X_{m_n,n}$

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

Attribute Vector		
$Y_{\Theta^{-1}(1),1}$	$\cdots$	$Y_{\Theta^{-1}(1),n}$
$\vdots$		$\vdots$
$Y_{\Theta^{-1}(m_n),1}$	$\cdots$	$Y_{\Theta^{-1}(m_n),n}$

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

Previous Works: Information-Theoretical Limits

[Shirani, Garg, and Erkip, ISIT '19]

$\mathbf{D}^{(1)}$

User ID	Attribute Vector		
1	$X_{1,1}$	$\cdots$	$X_{1,n}$
$\vdots$	$\vdots$		$\vdots$
m_n	$X_{m_n,1}$	$\cdots$	$X_{m_n,n}$

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

Attribute Vector		
$Y_{\Theta^{-1}(1),1}$	$\cdots$	$Y_{\Theta^{-1}(1),n}$
$\vdots$		$\vdots$
$Y_{\Theta^{-1}(m_n),1}$	$\cdots$	$Y_{\Theta^{-1}(m_n),n}$

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

Previous Works: Information-Theoretical Limits

[Shirani, Garg, and Erkip, ISIT '19]

$\mathbf{D}^{(1)}$

User ID	Attribute Vector		
1	$X_{1,1}$	$\cdots$	$X_{1,n}$
$\vdots$	$\vdots$		$\vdots$
m_n	$X_{m_n,1}$	$\cdots$	$X_{m_n,n}$

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

Attribute Vector		
$Y_{\Theta^{-1}(1),1}$	$\cdots$	$Y_{\Theta^{-1}(1),n}$
$\vdots$		$\vdots$
$Y_{\Theta^{-1}(m_n),1}$	$\cdots$	$Y_{\Theta^{-1}(m_n),n}$

- Databases as $m_n \times n$ random matrices: equal no. of labeled attributes (columns)
  - Matching rows $\sim f_{X^n,Y^n}$: Noise-only.
  - Non-matching rows $\sim f_{X^n} f_{Y^n}$:
- 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 a successful matching scheme $\hat{\Theta}$
  - Successful: $\lim_{n\to\infty} \Pr(\Theta(I) = \hat{\Theta}(I)) = 1$ where $I \sim U(1, m_n)$.

Previous Works: Information-Theoretical Limits

- Objective: Given $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$, find a successful matching scheme $\hat{\Theta}$
    - Successful: $\lim_{n\to\infty} \Pr(\Theta(I) = \hat{\Theta}(I)) = 1$ where $I \sim U(1, m_n)$.

- <u>Almost</u> all entries must be matched correctly.

Previous Works: Information-Theoretical Limits

- Objective: Given $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$, find a successful matching scheme $\hat{\Theta}$
  - Successful: $\lim_{n\to\infty} \Pr(\Theta(I) = \hat{\Theta}(I)) = 1$ 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 a successful matching scheme $\hat{\Theta}$
  - Successful: $\lim_{n\to\infty} \Pr(\Theta(I) = \hat{\Theta}(I)) = 1$ 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.
- Achievable Database Growth Rate: Rate R is achievable if given $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$ with growth rate R, there exists a successful matching scheme.

Previous Works: Information-Theoretical Limits

- Objective: Given $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$, find a successful matching scheme $\hat{\Theta}$
  - Successful: $\lim_{n\to\infty} \Pr(\Theta(I) = \hat{\Theta}(I)) = 1$ 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.
- Achievable Database Growth Rate: Rate R is achievable if given $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$ with growth rate R, there exists a successful matching scheme.
- Matching Capacity:

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

Theorem (Noise-Only Matching Capacity)

In the noise-only setting, the matching capacity is given by $C = I(X; Y)$.

Previous Works: Information-Theoretical Limits

1. **Random Deletions & Replications** [**Bakirtas** & Erkip, ISIT '21, Asilomar '22]
   - Underlying repetition distribution p_S over $\{0,\dots,s_{\max}\}$.

Previous Works: Information-Theoretical Limits

1. **Random Deletions & Replications** [**Bakirtas** & Erkip, ISIT '21, Asilomar '22]
   - Underlying repetition distribution p_S over $\{0, \dots, s_{\max}\}$.

Theorem (Repetition-Only Matching Capacity)

In the repetition-only setting, the matching capacity is equal to the erasure channel mutual information with erasure probability $p_S(0)$.

Previous Works: Information-Theoretical Limits

1. **Random Deletions & Replications** [**Bakirtas** & Erkip, ISIT '21, Asilomar '22]
   - Underlying repetition distribution p_S over $\{0, \dots, s_{\max}\}$.

Theorem (Repetition-Only Matching Capacity)

In the repetition-only setting, the matching capacity is equal to the erasure channel mutual information with erasure probability $p_S(0)$.

2. **Random Deletions & Replications** + **Noise** [**Bakirtas** & Erkip, ITW '22]
   - Seeds (already-matched row pairs) available.

Previous Works: Information-Theoretical Limits

1. **Random Deletions & Replications** [**Bakirtas** & Erkip, ISIT '21, Asilomar '22]
   - Underlying repetition distribution p_S over $\{0, \dots, s_{\max}\}$.

Theorem (Repetition-Only Matching Capacity)

In the repetition-only setting, the matching capacity is equal to the erasure channel mutual information with erasure probability $p_S(0)$.

2. **Random Deletions & Replications + Noise** [**Bakirtas** & Erkip, ITW '22]
   - Seeds (already-matched row pairs) available.

Theorem (Seeded Matching Capacity with Repetition + Noise)

Given a seed size $\Lambda_n = \Omega(\log m_n)$ the matching capacity is $C = I(X; Y^S, S)$.

System Model

- $\mathbf{D}^{(1)}$: $m_n \times n$ random matrix with entries $X_{i,j} \overset{i.i.d.}{\sim} p_X$.

System Model

- $\mathbf{D}^{(1)}$: $m_n \times n$ random matrix with entries $X_{i,j} \overset{i.i.d.}{\sim} p_X$.
- Database Growth Rate: $R = \lim_{n\to\infty} \frac{1}{n} \log m_n$
  - Assumption: $R > 0$. $(n \sim \log m_n)$

System Model

- $\mathbf{D}^{(1)}$: $m_n \times n$ random matrix with entries $X_{i,j} \overset{i.i.d.}{\sim} p_X$.
- Database Growth Rate: $R = \lim_{n\to\infty} \frac{1}{n} \log m_n$
  - Assumption: $R > 0$. $(n \sim \log m_n)$
  - Only interesting regime [Kunisky & Niles-Weed, 2022]

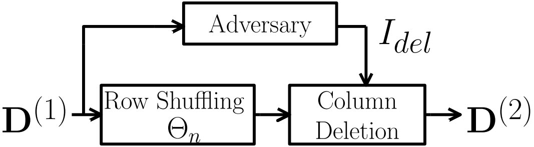
System Model

- $\mathbf{D}^{(1)}$: $m_n \times n$ random matrix with entries $X_{i,j} \overset{i.i.d.}{\sim} p_X$.
- Database Growth Rate: $R = \lim_{n\to\infty} \frac{1}{n} \log m_n$
  - Assumption: $R > 0$. $(n \sim \log m_n)$
  - Only interesting regime [Kunisky & Niles-Weed, 2022]
- Θ_n: Uniform permutation of $[m_n]$.

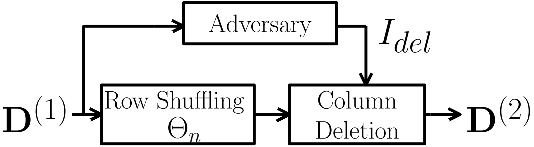
System Model

- $\mathbf{D}^{(1)}$: $m_n \times n$ random matrix with entries $X_{i,j} \overset{i.i.d.}{\sim} p_X$.
- Database Growth Rate: $R = \lim_{n\to\infty} \frac{1}{n} \log m_n$
  - Assumption: $R > 0$. $(n \sim \log m_n)$
  - Only interesting regime [Kunisky & Niles-Weed, 2022]
- Θ_n: Uniform permutation of $[m_n]$.
- Column deletion pattern: $I_{\text{del}} = \{i_1, i_2, \ldots, i_d\} \subseteq [n]$.
  - Chosen by an adversary after observing $\mathbf{D}^{(1)}$
  - $\delta \triangleq \frac{d}{n}$: Deletion budget
  - Identical deletion pattern across rows.

System Model: Continued

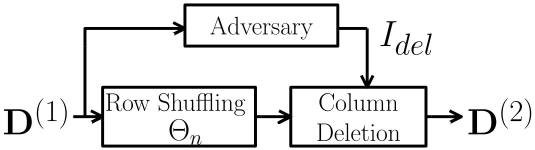


System Model: Continued



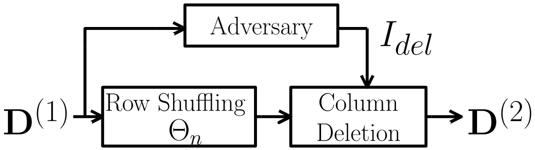
- $\mathbf{D}^{(2)}$: Obtained from $\mathbf{D}^{(1)}$ by
  1. Row shuffling by Θ_n.

System Model: Continued



- $\mathbf{D}^{(2)}$: Obtained from $\mathbf{D}^{(1)}$ by
  1. Row shuffling by Θ_n.
  2. Column deletion by I_{del}.
     - Delete the j^{th} column if $j \in I_{\text{del}}$.

System Model: Continued



- $\mathbf{D}^{(2)}$: Obtained from $\mathbf{D}^{(1)}$ by
  1. Row shuffling by Θ_n.
  2. Column deletion by I_{del}.
     - Delete the j^{th} column if $j \in I_{\text{del}}$.
- No noise on the entries.

System Model: Continued

- Achievable Database Growth Rate: Rate R is achievable if given $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$ with growth rate R, $\exists \hat{\Theta}_n$ such that:

$$\Pr(\forall I_{\text{del}} = (i_1, \dots, i_{n\delta}) \subseteq [n], \hat{\Theta}_n(J) = \Theta_n(J)) \xrightarrow{n\to\infty} 1,$$

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

- Adversarial Matching Capacity:

$$C^{\text{adv}}(\delta) \triangleq \sup\{R\colon\ R \text{ is achievable with deletion budget } \delta.\}$$

System Model: Continued

- Achievable Database Growth Rate: Rate R is achievable if given $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$ with growth rate R, $\exists \hat{\Theta}_n$ such that:

$$\Pr(\forall I_{\text{del}} = (i_1, \dots, i_{n\delta}) \subseteq [n], \hat{\Theta}_n(J) = \Theta_n(J)) \xrightarrow{n\to\infty} 1,$$

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

- Adversarial Matching Capacity:

$$C^{\text{adv}}(\delta) \triangleq \sup\{R\colon\ R \text{ is achievable with deletion budget } \delta.\}$$

- Goal: Given p_X and δ, characterize matching capacity $C^{\text{adv}}(\delta)$.

This Talk: Objectives

1. What is the adversarial matching capacity?

This Talk: Objectives

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

This Talk: Objectives

1. What is the adversarial matching capacity?
2. Can we devise matching schemes that achieve this matching capacity?
3. Can adversarial deletion offer better privacy than the random one?

Proposed Matching Scheme

- Exploit the identical deletion pattern across rows.

Proposed Matching Scheme

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

Proposed Matching Scheme

- Exploit the identical deletion pattern across rows.
  1. Find a permutation-invariant unique feature of the columns.
  2. By matching these features, infer the deletion pattern I_{del}.

Proposed Matching Scheme

- Exploit the identical deletion pattern across rows.
  1. Find a permutation-invariant unique feature of the columns.
  2. By matching these features, infer the deletion pattern I_{del}.
  3. Discard the deleted columns from $\mathbf{D}^{(1)}$.

Proposed Matching Scheme

- Exploit the identical deletion pattern across rows.
  1. Find a permutation-invariant unique feature of the columns.
  2. By matching these features, infer the deletion pattern I_{del}.
  3. Discard the deleted columns from $\mathbf{D}^{(1)}$.
  4. Perform rowwise exact sequence matching.
- We will use column histograms as the permutation-invariant feature.

Proposed Matching Scheme

- Exploit the identical deletion pattern across rows.
  1. Find a permutation-invariant unique feature of the columns.
  2. By matching these features, infer the deletion pattern I_{del}.
  3. Discard the deleted columns from $\mathbf{D}^{(1)}$.
  4. Perform rowwise exact sequence matching.
- 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 deletion pattern across rows.
  1. Find a permutation-invariant unique feature of the columns.
  2. By matching these features, infer the deletion pattern I_{del}.
  3. Discard the deleted columns from $\mathbf{D}^{(1)}$.
  4. Perform rowwise exact sequence matching.
- 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 deletion pattern.

Histogram-Based Repetition Detection: Example

$$\mathbf{D}^{(1)} = \begin{bmatrix} a & b & a & c & c & b \\ a & c & c & c & b & b \\ c & c & b & b & c & c \\ a & c & b & a & b & d \\ c & b & c & b & c & b \\ c & a & b & b & b & c \\ c & c & c & b & c & b \\ b & b & b & a & c & c \end{bmatrix} \rightarrow \mathbf{H}^{(1)} = \begin{bmatrix} 3 & 1 & 1 & 2 & 0 & 0 \\ 1 & 3 & 4 & 4 & 3 & 4 \\ 4 & 4 & 3 & 2 & 5 & 3 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Histogram-Based Repetition Detection: Example

$$\mathbf{D}^{(2)} = \begin{bmatrix} a & a & c & b \\ a & c & c & b \\ c & b & b & c \\ a & b & a & d \\ c & c & b & b \\ c & b & b & c \\ c & c & b & b \\ b & b & a & c \end{bmatrix} \longrightarrow \mathbf{H}^{(2)} = \begin{bmatrix} 3 & 1 & 2 & 0 \\ 1 & 4 & 4 & 4 \\ 4 & 3 & 2 & 3 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

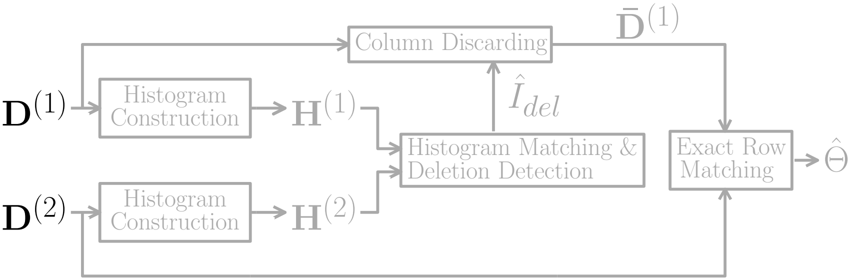
Histogram-Based Repetition Detection: Example

$$\mathbf{H}^{(1)} \qquad \mathbf{H}^{(2)}$$

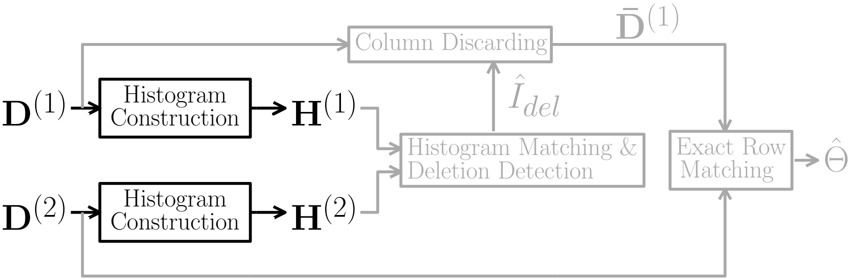
$$\boxed{\begin{matrix} 3 & 1 & 1 & 2 & 0 & 0 \\ 1 & 3 & 4 & 4 & 3 & 4 \\ 4 & 4 & 3 & 2 & 5 & 3 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{matrix}} \qquad \boxed{\begin{matrix} 3 & 1 & 2 & 0 \\ 1 & 4 & 4 & 4 \\ 4 & 3 & 2 & 3 \\ 0 & 0 & 0 & 1 \end{matrix}}$$

$$\hat{I}_{del} = [2 \;\; 5]$$

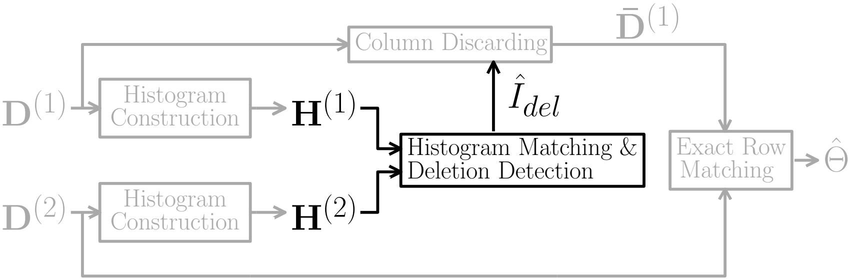
Matching Scheme



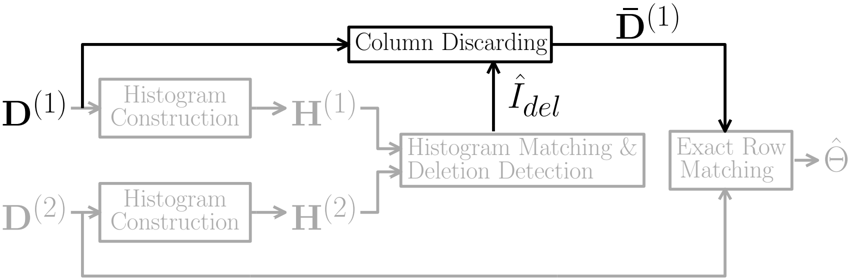
Matching Scheme



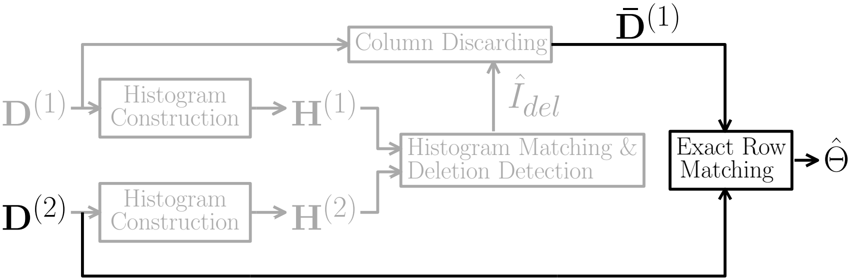
Matching Scheme



Matching Scheme



Matching Scheme



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 $m_n = \omega\left(n^{\frac{4}{|\mathfrak{X}|-1}}\right)$.

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 $m_n = \omega\left(n^{\frac{4}{|\mathfrak{X}|-1}}\right)$.

- 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 $m_n = \omega\left(n^{\frac{4}{|\mathfrak{X}|-1}}\right)$.

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

Asymptotic Uniqueness of The Histograms

Lemma

Let H_i denote the i^{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 $m_n = \omega\left(n^{\frac{4}{|\mathfrak{X}|-1}}\right)$.

- 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 $\mathbf{H}^{(2)}$.
- **Note:**
  - LLN: $H_i \approx H_j,\ \forall i,j$

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 $m_n = \omega\left(n^{\frac{4}{|\mathfrak{X}|-1}}\right)$.

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

- **Note:**
  - LLN: $H_i \approx H_j,\ \forall i,j$
  - Our Result: $H_i \approx H_j$, **BUT** $H_i \neq H_j$

Main Result: Adversarial Matching Capacity

Theorem (Adversarial Matching Capacity)

Consider a database distribution p_X and an adversary with a δ-*deletion budget*. Then, the adversarial matching capacity is

$$C^{\text{adv}}(\delta) = \begin{cases} D(\delta \| 1 - \hat{q}), & \text{if } \delta \leq 1 - \hat{q} \\ 0, & \text{if } \delta > 1 - \hat{q} \end{cases}$$

where $\hat{q} \triangleq \sum_{x \in \mathfrak{X}} p_X(x)^2$ and $D(.\|.)$ denotes the KL divergence between two Bernoulli distributions with given parameters.

Main Result: Adversarial vs. Random Deletion

- Adversarial Matching Capacity

$$C^{\mathsf{adv}}(\delta) = \begin{cases} D(\delta \| 1-\hat{q}), & \text{if } \delta \le 1-\hat{q} \\ 0, & \text{if } \delta > 1-\hat{q} \end{cases}$$

Main Result: Adversarial vs. Random Deletion

- Adversarial Matching Capacity

$$C^{\text{adv}}(\delta) = \begin{cases} D(\delta \| 1-\hat{q}), & \text{if } \delta \le 1-\hat{q} \\ 0, & \text{if } \delta > 1-\hat{q} \end{cases}$$

- Random Matching Capacity [**Bakirtas** & Erkip, Asilomar '22]

$$C^{\text{random}}(\delta) = (1-\delta)H(X)$$

Main Result: Adversarial vs. Random Deletion

- Adversarial Matching Capacity

$$C^{\text{adv}}(\delta) = \begin{cases} D(\delta \| 1-\hat{q}), & \text{if } \delta \leq 1-\hat{q} \\ 0, & \text{if } \delta > 1-\hat{q} \end{cases}$$

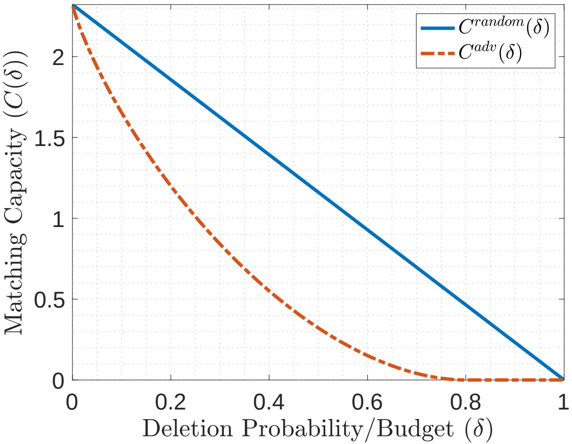
- Random Matching Capacity [**Bakirtas** & Erkip, Asilomar '22]

$$C^{\text{random}}(\delta) = (1-\delta)H(X)$$

  - Strictly positive!

Adversarial vs. Random Deletion: Example

$X \sim \mathrm{Unif}(\mathfrak{X})$, $\mathfrak{X} = [5]$. $1 - \hat{q} = 0.8$.



Conclusion

- Database Matching $\Leftrightarrow$ Channel Decoding

Conclusion

- Database Matching $\Leftrightarrow$ Channel Decoding
- Histograms help us infer the deletion pattern.

Conclusion

- Database Matching $\Leftrightarrow$ Channel Decoding
- Histograms help us infer the deletion pattern.
- Complete characterization of the adversarial matching capacity.

Conclusion

- Database Matching $\Leftrightarrow$ Channel Decoding
- Histograms help us infer the deletion pattern.
- Complete characterization of the adversarial matching capacity.
- Adversarial deletions offer better privacy, compared to random deletions.

Conclusion

- Database Matching $\Leftrightarrow$ Channel Decoding
- Histograms help us infer the deletion pattern.
- Complete characterization of the adversarial matching capacity.
- Adversarial deletions offer better privacy, compared to random deletions.
- Ongoing Work: Database matching with adversarial noise, distribution-agnostic database matching.

Thank you! Q&A?

Database Matching Under Adversarial Column Deletions

Serhat Bakirtas, Elza Erkip

serhat.bakirtas@nyu.edu

NYU | TANDON SCHOOL OF ENGINEERING

NYU WIRELESS