Distribution-Agnostic Database De-Anonymization Under Synchronization Errors

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

***Abstract*—Recently, there has been an increased scientific interest in the de-anonymization of users in anonymized databases containing user-level microdata via multifarious matching strategies utilizing publicly-available correlated data. Existing literature has either emphasized practical aspects where underlying data distribution is not required, with limited or no theoretical guarantees, or theoretical aspects with the assumptions of complete availability of underlying distributions. In this work, we take a step towards reconciling these two lines of work, by providing theoretical guarantees for the de-anonymization of random correlated databases, without prior knowledge of data distribution. Motivated by time-indexed microdata, we consider database de-anonymization under both synchronization errors (column repetitions) and obfuscation (noise). By modifying the existing replica detection algorithm to accommodate for the unknown underlying distribution, proposing a new seeded deletion detection algorithm, and employing statistical and information-theoretic tools, we derive sufficient conditions on the database growth rate for successful matching. Our findings demonstrate that a double-logarithmic seed size relative to row size ensures successful deletion detection. More importantly, we show that the derived sufficient conditions are the same as in the distribution-aware setting, negating any asymptotic loss of performance due to unknown underlying distributions.**

I. Introduction

With the accelerating growth of smart devices and applications, there has been a considerable collection of user-level microdata in private companies' and public institutions' possession which is often shared and/or sold. Although this data transfer is performed after removing the explicit user identifiers, a.k.a. *anonymization*, and coarsening of the data through noise, a.k.a. *obfuscation*, there is a growing concern from the scientific community [1] about the privacy implications. These concerns were further validated by the success of a series of practical attacks on real data by researchers [2]–[6]. In the light of these successful experimental work, recently there has been an increasing effort on the information-theoretic and statistical foundations of *database de-anonymization*, a.k.a. *database alignment/matching/recovery* [7]–[16].

More recently, we have focused on the database de-anonymization problem under synchronization errors. In [13], we investigated the matching of Markov databases under synchronization errors, with no subsequent obfuscation/noise. We showed that the synchronization errors could be detected through a histogram-based detection scheme. Furthermore, we found the noiseless matching capacity to be equal to the erasure bound where locations of deletions and replications are known a-priori. More relevantly, in [14], we considered the de-anonymization of databases under noisy synchronization errors. We proposed a noisy replica detection algorithm and a seeded deletion detection algorithm to detect synchronization errors. We proposed a joint-typicality-based matching algorithm and derived achievability results, which we subsequently showed to be tight, given a seed size logarithmic with the row size of the database. Then in [15], we improved this sufficient seed size to one double logarithmic with the row size. Albeit successful in deriving detecting and matching results, in these works, the availability of information on the underlying distributions was assumed and the proposed algorithms were tailored for these known distributions.

This work is supported in part by National Science Foundation grants 1815821, 2148293, and 2003182, and NYU WIRELESS Industrial Affiliates.

Motivated by most practical settings where the underlying distributions are not readily available, but only could be estimated from the available data, in this paper, we investigate the de-anonymization problem without any a priori knowledge of the underlying distributions. We focus on a noisy random column repetition model borrowed from [14], as illustrated in Figure 1. We modify the noisy replica detection algorithm proposed in [14] so that it still works in the novel distribution-agnostic setting. Then we propose a novel outlier-detection-based deletion detection algorithm and show that when seeds whose size grows double logarithmic with the number of users (rows) in the database, the underlying deletion pattern could be inferred. Finally, through a typicality-based de-anonymization algorithm that relies on the estimated distributions, we show that database de-anonymization could be performed with no asymptotic loss of performance compared to when all the information on the distributions is available a priori.

The structure of the rest of this paper is as follows: Section II introduces the formal statement of the problem. Section III contains our proposed algorithms, states our main result, and contains its proof. Finally, Section IV consists of the concluding remarks.

Notation: We denote a matrix $\mathbf{D}$ with bold capital letters, and its $(i,j)^{\text{th}}$ element with $D_{i,j}$. A set is denoted by a calligraphic letter, *e.g.*, $\mathcal{X}$. $[n]$ denotes the set of integers $\{1,\dots,n\}$. Asymptotic order relations are used as defined in [17, Chapter 3]. All logarithms are base 2. $H(.)$ denotes the Shannon entropy [18, Chapter 2].

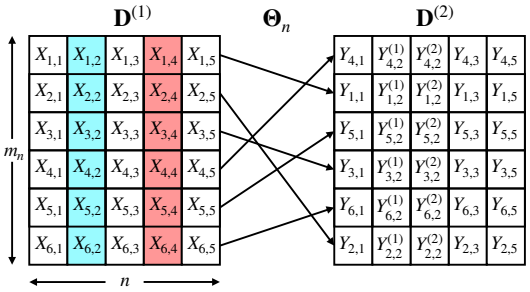

Fig. 1: An illustrative example of database matching under column repetitions. The column coloured in red is repeated zero times, *i.e.,* deleted, whereas the column coloured in blue is repeated twice, *i.e.,* replicated. $Y^{(1)}_{i,2}$ and $Y^{(2)}_{i,2}$ denote noisy copies/replicas of $X_{i,2}$. Our goal is to estimate the correct row permutation $\Theta_n = \left(\begin{smallmatrix}1&2&3&4&5&6\\2&6&4&1&3&5\end{smallmatrix}\right)$, by matching the rows of $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$.

II. Problem Formulation

We use the following definitions, most of which are borrowed from [14] to formalize our problem.

Definition 1. (Anonymized Database) An (m_n, n, p_X) *anonymized database* $\mathbf{D} = \{X_{i,j} \in \mathfrak{X}\}$ is a randomly generated $m_n \times n$ matrix with $X_{i,j} \overset{\text{i.i.d.}}{\sim} p_X$, where p_X has a finite discrete support $\mathfrak{X} = \{1, \ldots, |\mathfrak{X}|\}$.

Definition 2. (Column Repetition Pattern) The *column repetition pattern* $S^n = \{S_1, S_2, ..., S_n\}$ is a random vector with $S_i \overset{\text{i.i.d.}}{\sim} p_S$, where p_S has a finite integer support $\{0, \ldots, s_{\max}\}$.

Remark 1. We note the fact that $X_{i,j}$ and S_i are *i.i.d.* can be checked through the Markov order estimation algorithm of [19] with a probability of error vanishing in n. Thus from now on, we assume that the *i.i.d.* nature of $X_{i,j}$ and S_i is known, while the distributions p_X and p_S are not.

Remark 2. Since $|\mathfrak{X}|$ and $s_{\max}$ do not depend on n, they can easily be estimated with a probability of error vanishing in n. Therefore, we will assume that $|\mathfrak{X}|$ and $s_{\max}$ are known.

Definition 3. (Anonymization Function) The *anonymization function* Θ_n is a uniformly-drawn permutation of $[m_n]$.

Definition 4. (Labeled Correlated Database) Let $\mathbf{D}^{(1)}$, S^n and Θ_n be a mutually-independent (m_n, n, p_X) anonymized database, repetition pattern and anonymization function triplet. Let $p_{Y|X}$ be a conditional probability distribution with both X and Y taking values from $\mathfrak{X}$. Given $\mathbf{D}^{(1)}$, S^n, Θ_n and $p_{Y|X}$, $\mathbf{D}^{(2)}$ is called the *labeled correlated database* if the respective $(i,j)^{\text{th}}$ entries $X_{i,j}$ and $Y_{i,j}$ of $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$ have the following relation:

$$Y_{\Theta_n(i),j} = \begin{cases} E, & \text{if } S_j = 0 \\ Z^{S_j} & \text{if } S_j \geq 1 \end{cases} \quad \forall i \in [m_n], \forall j \in [n] \tag{1}$$

where Z^{S_j} is a row vector consisting of S_j noisy replicas of $X_{i,j}$ with the following conditional probability distribution

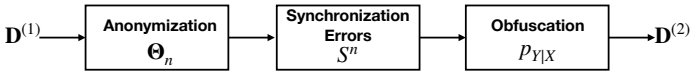

Fig. 2: Relation between the unlabeled database $\mathbf{D}^{(1)}$ and the labeled noisy repeated one, $\mathbf{D}^{(2)}$.

$$\Pr\left(Z^{S_j} = z^{S_j} | X_{i,j} = x\right) = \prod_{l=1}^{S_j} p_{Y|X}\left(z_l | x\right) \tag{2}$$

where $z^{S_j} = z_1, \ldots, z_{S_j}$ and $Y_{\Theta_n(i),j} = E$ corresponds to $Y_{\Theta_n(i),j}$ being the empty string.

Note that S_j indicates the times the j^{th} column of $\mathbf{D}^{(1)}$ is repeated. When $S_j = 0$, the j^{th} column of $\mathbf{D}^{(1)}$ is said to be *deleted* and when $S_j > 1$, the j^{th} column of $\mathbf{D}^{(1)}$ is said to be *replicated*.

The i^{th} row X_i of $\mathbf{D}^{(1)}$ and the $\Theta_n(i)^{\text{th}}$ row $Y_{\Theta_n(i)}$ of $\mathbf{D}^{(2)}$ are called *matching rows*.

The relationship between $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$, as described in Definition 4, is illustrated in Figure 2.

Remark 3. In this work, we assume a memoryless noise model, so that the conditional independence of the noisy replicas stated in (2) is known, whereas the noise distribution $p_{Y|X}$ is not.

As often done in both the graph matching [20] and the database matching [14] literatures, we will assume the availability of a set of already-matched row pairs called *seeds*, to be used in the detection of the underlying repetition pattern.

Definition 5. (Seeds) Given a pair of anonymized and labeled correlated databases $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$, a *seed* is a correctly-matched row pair with the same underlying repetition pattern. A *batch of* Λ_n *seeds* is a pair of seed matrices of respective sizes $\Lambda_n \times n$ and $\Lambda_n \times \sum_{j=1}^{n} S_j$.

For the sake of notational brevity, we assume that the seed matrices $\mathbf{G}^{(1)}$ and $\mathbf{G}^{(2)}$) are not submatrices of $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$. Throughout this work, we will assume a seed size $\Lambda_n = \omega(\log n) = \omega(\log\log m_n)$ which is double-logarithmic with the number of users m_n.

As done in [8], [12]–[14], [16], we utilize the database growth rate, defined below, as the main performance metric.

Definition 6. (Database Growth Rate) The *database growth rate* R of an (m_n, n, p_X) anonymized database is defined as

$$R = \lim_{n\to\infty} \frac{1}{n} \log m_n \tag{3}$$

Similar to [8], [12]–[14], in this paper, our goal is to characterize the supremum of the achievable database growth rates allowing the *almost-perfect* recovery of the anonymization function Θ_n. However, unlike [8], [12]–[14], we consider the case *when the underlying distributions p_X, $p_{Y|X}$ and p_S are not*

provided a priori. More formally, "almost-perfect recovery" corresponds to the construction of the estimate $\hat{\Theta}_n$ such that

$$\lim_{n\to\infty} \Pr\left(\Theta_n(J) \neq \hat{\Theta}_n(J)\right) \to 0 \tag{4}$$

where $J \sim \text{Unif}([m_n])$.

III. De-Anonymization Algorithm and Achievability

In this section, we present our main result on the achievable database growth rates when no a-priori information is provided on p_X, $p_{Y|X}$, and p_S. We state the main result in Theorem 1 and prove it.

Theorem 1. ***(Main Result)*** *Consider an anonymized and labeled correlated database pair, with underlying database distributions $p_{X,Y}$ and a column repetition distribution p_S which are assumed to be not known a-priori. Given a seed size $\Lambda_n = \omega(\log n)$, any database growth rate R satisfying*

$$R < I(X;Y^S|S) \tag{5}$$

is achievable where $S \sim p_S$, $X \sim p_X$ and $Y^S = Y_1,\ldots,Y_S$ with $Y_i|X \overset{i.i.d.}{\sim} p_{Y|X}$.

In order to demonstrate the tightness of the achievability result stated in Theorem 1, we now compare it to the distribution-aware results derived in [14, Theorem 1].

Theorem 2. ***(Converse of [14, Theorem 1])*** *Consider an anonymized and labeled correlated database pair, with underlying joint database distributions $p_{X,Y}$ and a column repetition distribution p_S. Then, a necessary condition for the existence of a successful de-anonymization scheme is:*

$$R \leq I(X;Y^S|S) \tag{6}$$

Theorems 1 and 2 imply that given a seed size $\Lambda_n = \omega(\log n) = \omega(\log\log m_n)$ we can perform matching as if we knew the underlying distribution $p_{X,Y}$ and the actual column repetition pattern S^n a-priori. Hence in the asymptotic regime, not knowing the distributions causes no loss in the matching capacity.

The rest of this section is on the proof of Theorem 1. In Section III-A, we present our detection of noisy replicas algorithm and prove its asymptotic performance. Then in Section III-B, we propose a seeded deletion algorithm and derive a sufficient seed size that guarantees its asymptotic performance. Finally in Section III-C, we present our de-anonymization algorithm.

A. Noisy Replica Detection

Similar to [14], we use the running Hamming distances between the consecutive columns $C_j^{(2)}$ and $C_{j+1}^{(2)}$ of $\mathbf{D}^{(2)}$, denoted by H_j, $j \in [K_n - 1]$, where $K_n \triangleq \sum_{j=1}^n S_j$ as a permutation-invariant future of the labeled correlated database. More formally,

$$H_j \triangleq \sum_{t=1}^{m_n} \mathbb{1}_{[D^{(2)}_{t,j+1} \neq D^{(2)}_{t,j+1}]}, \qquad \forall j \in [K_n - 1] \tag{7}$$

Algorithm 1: Noisy Replica Detection Algorithm

Input : $(\mathbf{D}, m_n, K_n)$
Output: isReplica
$H \leftarrow$ RunningHammingDist($\mathbf{D}$); /* Eq. (7) */
$(\hat{p}_0, \hat{p}_1) \leftarrow$EstimateParams($H$); /* See [21] */
$\tau \leftarrow \frac{\hat{p}_0+\hat{p}_1}{2}$; /* Threshold */
isReplica $\leftarrow \varnothing$;
for $j = 1$ **to** $K_n - 1$ **do**
  if $H[j] \leq m_n\tau$ **then**
    isReplica$[j] \leftarrow$ TRUE;
  else
    isReplica$[j] \leftarrow$ FALSE;
  end
end

We first note that

$$H_j \sim \begin{cases} \text{Binom}(m_n, p_0), & \text{if } C_j^{(2)} \perp\!\!\!\perp C_{j+1}^{(2)} \\ \text{Binom}(m_n, p_1), & \text{otherwise} \end{cases} \tag{8}$$

where

$$p_0 \triangleq 1 - \sum_{y\in\mathfrak{X}} p_Y(y)^2 \tag{9}$$

$$p_1 \triangleq 1 - \sum_{x\in\mathfrak{X}} p_X(x) \sum_{y\in\mathfrak{X}} p_{Y|X}(y|x)^2 \tag{10}$$

From [16, Lemma 1], we know that as long as the databases are correlated, *i.e.*, $p_{X,Y} \neq p_X p_Y$, we have $p_0 > p_1$ for any $p_{X,Y}$. Thus, as long as $p_{X,Y} \neq p_X p_Y$, replicas can be detected based on the Hamming distances H_j similar to [14], [16]. However, the algorithm in [14] depends on the chose of a threshold that depends on $p_{X,Y}$. In Algorithm 1, we propose the following modification: We first construct the estimates $\hat{p}_0$ and $\hat{p}_1$ for the respective parameters p_0 and p_1 through the moment estimator proposed by Blischke in [21]. Note that we can use this estimator because the Binomial mixture is guaranteed to have two distinct components. More formally, the distribution of H_j conditioned on S^n is given by

$$\Pr(H_j = h|S^n) = \binom{m_n}{h}[\alpha p_0^h(1-p_0)^{m_n-h} + (1-\alpha)p_1^h(1-p_1)^{m_n-h}] \tag{11}$$

for $h = 0,\ldots,m_n$ where the mixing parameter α is given by

$$\alpha = \frac{1}{K_n - 1}\left(n - \sum_{j=1}^n \mathbb{1}_{[S_j=0]}\right) \tag{12}$$

It can easily be verified that as $n \to \infty$, $\alpha \xrightarrow{p} \frac{1-\delta}{\mathbb{E}[S]}$. Hence it is bounded away from both 0 and 1, suggesting that Algorithm 1 can be used to detect the replicas. More formally, for any $\varepsilon > 0$.

$$\lim_{n\to\infty} \Pr\left(\left|\alpha - \frac{1-\delta}{\mathbb{E}[S]}\right| > \varepsilon\right) = 0. \tag{13}$$

The following lemma states that this algorithm has a vanishing error probability.

Lemma 1. ***(Noisy Replica Detection)*** *Algorithm 1 has a vanishing probability of replica detection error, as long as* $m_n = \omega(\log n)$.

Proof. The estimator proposed in [21] works as follows: Define the k^{th} sample factorial moment F_k as

$$F_k \triangleq \frac{1}{K_n} \sum_{j=1}^{K_n} \prod_{i=0}^{k-1} \frac{H_j - i}{m_n - i}, \quad \forall k \in [m_n] \tag{14}$$

and let

$$A \triangleq \frac{F_3 - F_1 F_2}{F_2 - F_1^2} \tag{15}$$

Then the respective estimators $\hat{p}_0$ and $\hat{p}_1$ for p_0 and p_1 can be constructed as:

$$\hat{p}_0 = \frac{A + \sqrt{A^2 - 4AF_1 + 4F_2}}{2} \tag{16}$$

$$\hat{p}_1 = \frac{A - \sqrt{A^2 - 4AF_1 + 4F_2}}{2} \tag{17}$$

From [21], we get $\hat{p}_i \xrightarrow{p} p_i$ and in turn $\tau \xrightarrow{p} \frac{p_0+p_1}{2}$. Thus for large n, τ is bounded away from p_0 and p_1. Finally from [15, Lemma 1], we have

$$\lim_{n\to\infty} \Pr(\text{Noisy replica detection error}) = 0 \tag{18}$$

□

Note that the condition in Lemma 1 is automatically satisfied since m_n is exponential in n (Definition 6). Finally, we stress that as opposed to deletion detection, discussed in Section III-B, no seeds are necessary for replica detection.

B. Deletion Detection

In this section, we propose a deletion detection algorithm that utilizes the seeds. Since the replica detection algorithm of Section III-A (Algorithm 1) has a vanishing probability of error, for notational simplicity we will focus on a deletion-only setting throughout this subsection. Let $\mathbf{G}^{(1)}$ and $\mathbf{G}^{(2)}$ be the seed matrices with respective sizes $\Lambda_n \times n$ and $\Lambda_n \times \tilde{K}_n$, and denote the j^{th} column of $\mathbf{G}^{(r)}$ with $G_j^{(r)}$, $r = 1,2$ where $\tilde{K}_n \triangleq \sum_{j=1}^{n} \mathbb{1}_{[S_j \neq 0]}$. Furthermore, for the sake of brevity, let $L_{i,j}$ denote the Hamming distance between $G_i^{(1)}$ and $G_j^{(2)}$ for $(i,j) \in [n] \times [\tilde{K}_n]$. More formally, let

$$L_{i,j} \triangleq \sum_{t=1}^{\Lambda_n} \mathbb{1}_{[G_{t,i}^{(1)} \neq G_{t,j}^{(2)}]} \tag{19}$$

Observe that

$$L_{i,j} \sim \begin{cases} \text{Binom}(\Lambda_n, q_0), & G_i^{(1)} \perp\!\!\!\perp G_j^{(2)} \\ \text{Binom}(\Lambda_n, q_1), & \text{otherwise} \end{cases} \tag{20}$$

where

$$q_0 = 1 - \sum_{x \in \mathfrak{X}} p_X(x) p_Y(x) \tag{21}$$

$$q_1 = 1 - \sum_{x \in \mathfrak{X}} p_{X,Y}(x,x) \tag{22}$$

Thus, we have a problem seemingly similar to the one in Section III-A. However, we cannot utilize similar tools because of the following: *i)* Recall that the two components of the Binomial mixture discussed in Section III-A were distinct for any underlying joint distribution $p_{X,Y}$ as long as the databases are correlated, *i.e.,* $p_{X,Y} \neq p_X p_Y$. Unfortunately, the same idea does not automatically work here as demonstrated by the following example: Suppose $X_{i,j} \sim \text{Unif}(\mathfrak{X})$, and the transition matrix $\mathbf{P}$ associated with $p_{Y|X}$ has unit trace. Then,

$$q_0 - q_1 = \sum_{x \in \mathfrak{X}} p_{X,Y}(x,x) - p_X(x) p_Y(x) \tag{23}$$

$$= \frac{1}{|\mathfrak{X}|} \sum_{x \in \mathfrak{X}} p_{Y|X}(x|x) - p_Y(x) \tag{24}$$

$$= \frac{1}{|\mathfrak{X}|} (\text{tr}(\mathbf{P}) - 1) \tag{25}$$

$$= 0 \tag{26}$$

In [14], we overcame this problem using the following modification: Based on $p_{X,Y}$, we picked a bijective remapping $\Phi \in \mathfrak{S}(\mathfrak{X})$ and applied it to all the entries of $\mathbf{G}^{(2)}$ before computing the Hamming distances $L_{i,j}$, where $\mathfrak{S}(\mathfrak{X})$ denotes the symmetry group of $\mathfrak{X}$. Denoting the resulting version of the Hamming distance $L_{i,j}$ by $L_{i,j}(\Phi)$, we proved in [14, Lemma 2] that there as long as $p_{X,Y} \neq p_X p_Y$, there exists $\Phi \in \mathfrak{S}(\mathfrak{X})$ such that the Binomial mixture distribution associated with $L_{i,j}(\Phi)$ has two distinct components with respective success parameters $q_0(\Phi)$ and $q_1(\Phi)$. In other words, we have

$$L_{i,j}(\Phi) \sim \begin{cases} \text{Binom}(m_n, q_0(\Phi)), & G_i^{(1)} \perp\!\!\!\perp G_j^{(2)} \\ \text{Binom}(m_n, q_1(\Phi)), & \text{otherwise} \end{cases} \tag{27}$$

and $q_0(\Phi) \neq q_1(\Phi)$. We will such Φ a *useful remapping*.

ii) In the known-distribution setting, we chose the useful remapping Φ and threshold τ_n for Hamming distances based on $p_{X,Y}$. In Section III-A, we solved the distribution-agnostic case via parameter estimation in Binomial mixtures. However, the same approach does not work here. Suppose the j^{th} retained column $G_j^{(2)}$ of $\mathbf{G}^{(2)}$ is correlated with $G_{r_j}^{(1)}$. Then the j^{th} column of $\mathbf{L}(\Phi)$ will have a $\text{Binom}(\Lambda_n, q_1(\Phi))$ component in the $r_j{}^{\text{th}}$ row, whereas the remaining $n-1$ rows will contain $\text{Binom}(\Lambda_n, q_0(\Phi))$ components, as described in (27) and illustrated in Figure 3. Hence, it can be seen that the mixture parameter β of this Binomial mixture distribution approaches 1 since

$$\beta = \frac{(n-1)\tilde{K}_n}{n\tilde{K}_n} = 1 - \frac{1}{n} \tag{28}$$

This imbalance prevents us from performing a parameter estimation as done in Algorithm 1.

We propose to exploit the aforementioned observation that for a useful mapping Φ, in each column of $\mathbf{L}(\Phi)$, there is exactly one element with a different underlying distribution, while the remaining $n-1$ entries are *i.i.d.*, rendering this entry an *outlier*. Note that $L_{i,j}(\Phi)$ being an outlier corresponds to $G_i^{(1)}$ and $G_j^{(2)}$ being correlated, and in turn $S_i \neq 0$. On the other hand, we stress that the lack of outliers in any given column of

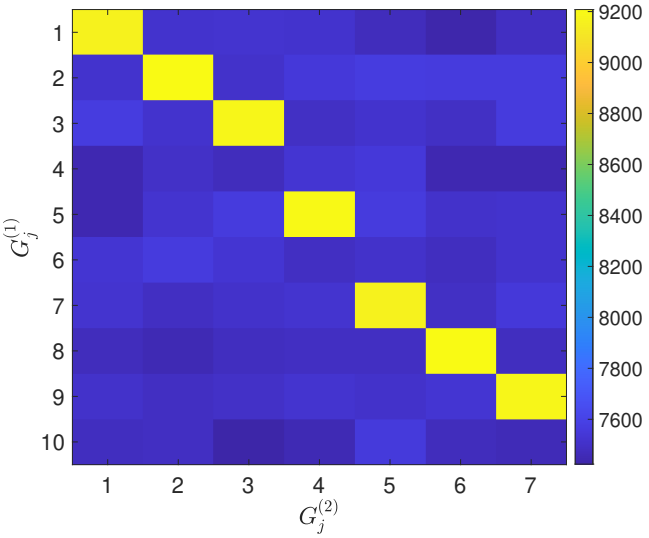

Fig. 3: Hamming distances between the columns of $\mathbf{G}^{(1)}$ and $\mathbf{G}^{(2)}$ with $n = 10$, $\tilde{K}_n = 7$ and $\Lambda_n = 10^4$. The $(i,j)^{\text{th}}$ element corresponds to $L_{i,j}$, with the colorbar indicating the approximate values. It can be seen that there are no outliers in the 4^{th}, 6^{th} and 10^{th} rows. Hence, it can be inferred that $I_{\text{del}} = (4,6,10)$.

Algorithm 2: Seeded Deletion Detection Algorithm

Input : $(\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \Lambda_n, n, \tilde{K}_n, \mathfrak{X})$
Output: retentionIdx
$\mathfrak{S}(\mathfrak{X}) \leftarrow$ SymmetryGroup($\mathfrak{X}$);
$\hat{\tau}_n \leftarrow 2\Lambda_n^{2/3}(\log n)^{1/3}$; /* Threshold */
for $s \leftarrow 1$ **to** $|\mathfrak{X}|!$ **do**
  retentionIdx$\leftarrow \varnothing$;
  $\Phi \leftarrow \mathfrak{S}(\mathfrak{X})[s]$; /* Pick a remapping. */
  $\mathbf{L}(\Phi) \leftarrow$ HammDist($\mathbf{G}^{(1)}, \mathbf{G}^{(2)}$); /* Eq. (19) */
  $\mu(\Phi) \leftarrow$ SampleMean($\mathbf{L}(\Phi)$); /* Eq. (29) */
  $\mathbf{M}(\Phi) \leftarrow |\mathbf{L}(\Phi) - \mu(\Phi)|$;
  for $j \leftarrow 1$ **to** $\tilde{K}_n$ **do**
    count $\leftarrow 0$;
    for $i \leftarrow 1$ **to** n **do**
      if $\mathbf{M}(\Phi)[i][j] \leq \hat{\tau}_n$ **then**
        retentionIdx$\leftarrow$ retentionIdx $\cup\{i\}$;
        count $\leftarrow$ count $+ 1$;
      end
    end
    /* count = 0: no outliers (Φ is useless). */
    /* count > 1: misdetection. */
    if count > 1 **then**
      return ERROR
    else
      if count $= 0$ **then**
        Skip to next Φ;
      end
    end
  end
  return $\hat{I}_R$;
end

$\mathbf{L}(\Phi)$ implies that Φ is useless. Thus, it can easily be seen that Algorithm 2 is capable of deciding whether a given remapping is useful or not. In fact, the algorithm sweeps over all elements of $\mathfrak{S}(\mathfrak{X})$ until we encounter a useful one.

To detect the outliers in $\mathbf{L}(\Phi)$, we propose to use the distances of $L_{i,j}(\Phi)$ to the sample mean $\mu(\Phi)$ of $\mathbf{L}(\Phi)$ where

$$\mu(\Phi) \triangleq \frac{1}{n\tilde{K}_n} \sum_{i=1}^{n} \sum_{j=1}^{\tilde{K}_n} L_{i,j}(\Phi) \tag{29}$$

As given in Algorithm 2, if these distances are lower than $\hat{\tau}_n$, we detect retention *i.e.,* non-deletion.

Note that this step is equivalent to utilizing Z-scores (also known as standard scores), a well-studied concept in statistical outlier detection [22], where the distances to the sample mean are also divided by the sample standard deviation. In this section, for the sake of brevity, we will avoid such division.

The following lemma states that for sufficient seed size, $\Lambda_n = \omega(\log n) = \omega(\log\log m_n)$, Algorithm 2 works correctly with high probability.

Lemma 2. ***(Deletion Detection)*** *Let $I_R = \{j \in [n] : S_j \neq 0\}$ be the true retention index set and $\hat{I}_R$ be its estimate output by Algorithm 2. Then for any seed size $\Lambda_n = \omega(\log n)$, we have*

$$\lim_{n\to\infty} \Pr(\hat{I}_R = I_R) = 1 \tag{30}$$

Proof. For now, suppose that Φ is a useful remapping. Start by observing that using Chebyshev's inequality [23, Theorem 4.2] it is straightforward to prove that for any $\varepsilon_n > 0$

$$\gamma \triangleq \Pr(|\mu(\Phi) - \Lambda_n q_0(\Phi)| > \Lambda_n \varepsilon_n) \leq O\left(\frac{1}{K_n n \Lambda_n \varepsilon_n}\right) \tag{31}$$

Let $I_R = \{r_1, \ldots, r_{\tilde{K}_n}\}$ and note that $L_{i,j} \sim \text{Binom}(\Lambda_n, q_1(\Phi))$. Thus, from the Chernoff bound [24, Lemma 4.7.2] we get

$$\begin{aligned} \beta_{r_j,j} &\triangleq \Pr(|L_{r_j,j}(\Phi) - \Lambda_n q_1(\Phi)| \geq \varepsilon_n \Lambda_n) && (32)\\ &\leq 2^{-\Lambda_n D(q_1(\Phi) - \varepsilon_n \| q_1(\Phi))} \\ &\quad + 2^{-\Lambda_n D(1-q_1(\Phi) - \varepsilon_n \| 1-q_1(\Phi))} && (33) \end{aligned}$$

where $D(p\|q)$ denotes the relative entropy [18, Chapter 2.3] (in bits) between two Bernoulli distributions with respective parameters p and q.

Now, for notational brevity, let

$$\begin{aligned} f(\varepsilon) &\triangleq D(q - \varepsilon \| q) && (34)\\ g(\varepsilon) &\triangleq D(1 - q - \varepsilon \| 1 - q) && (35) \end{aligned}$$

Then, one can simply verify the following

$$f'(\varepsilon) = \log\frac{q}{1-q} - \log\frac{q-\varepsilon}{1-q-\varepsilon} \tag{36}$$

$$f''(\varepsilon) = \frac{1}{\log e}\left[\frac{1}{q-\varepsilon} + \frac{1}{1-q+\varepsilon}\right] \tag{37}$$

$$g'(\varepsilon) = \log\frac{1-q}{q} - \log\frac{1-q-\varepsilon}{q+\varepsilon} \tag{38}$$

$$g''(\varepsilon) = \frac{1}{\log e}\left[\frac{1}{1-q-\varepsilon} + \frac{1}{q+\varepsilon}\right] \tag{39}$$

Observing that

$$f(0) = f'(0) = 0 \tag{40}$$

$$g(0) = g'(0) = 0 \tag{41}$$

and performing second-order MacLaurin Series expansions on f and g, we get for any $\varepsilon < 1$

$$f(\varepsilon) = c(q)\varepsilon^2 + O(\varepsilon^3) \tag{42}$$

$$g(\varepsilon) = c(q)\varepsilon^2 + O(\varepsilon^3) \tag{43}$$

where

$$c(q) \triangleq \frac{1}{2\log e}\left[\frac{1}{q} + \frac{1}{1-q}\right] \tag{44}$$

Let $\Lambda_n = \Gamma_n \log n$ and $\varepsilon_n = \Gamma_n^{-1/3}$ and pick the threshold as $\hat{\tau}_n = 2\Lambda_n\varepsilon_n$. Observe that since $\Gamma_n = \omega_n(1)$, we get

$$\hat{\tau}_n = 2\Lambda_n\varepsilon_n = o_n(\Lambda_n) \tag{45}$$

$$\Lambda_n\varepsilon_n^2 = \Gamma_n^{1/3}\log n = \omega_n(\log n) \tag{46}$$

Then, we have

$$\beta_{r_j,j} \le 2^{1-\Lambda_n(c(q_1(\Phi))\varepsilon_n^2 + O(\varepsilon_n^3))} \tag{47}$$

$$= 2^{1-c(q_1(\Phi))\Gamma_n^{1/3}\log n + O(\varepsilon_n^3))} \tag{48}$$

Note that with probability at least $1-\gamma-\beta_{r_j,j}$ we have

$$|\mu(\Phi) - \Lambda_n q_0(\Phi)| \le \Lambda_n\varepsilon_n \tag{49}$$

$$|L_{r_j,j}(\Phi) - \Lambda_n q_1(\Phi)| \ge \Lambda_n\varepsilon_n \tag{50}$$

From the triangle inequality, we have

$$|L_{r_j,j}(\Phi) - \mu(\Phi)| \ge \Lambda_n(|q_1(\Phi) - q_0(\Phi)| - 2\varepsilon_n) \tag{51}$$

$$\ge \hat{\tau}_n \tag{52}$$

for large n. Therefore, from the union bound we have

$$\Pr(\exists j \in [\tilde{K}_n] : |L_{r_j,j} - \mu(\Phi)| \le \hat{\tau}_n) \tag{53}$$

$$\le \gamma + \sum_{j=1}^{\tilde{K}_n} \beta_{r_j,j} \tag{54}$$

$$= \gamma + 2^{\log\tilde{K}_n - 1 - c(q_1(\Phi))\Gamma_n^{1/3}\log n + O(\varepsilon_n^3))} \tag{55}$$

Since $\tilde{K}_n \le n$ and $\Lambda_n = \omega_n(\log n)$, we have

$$\lim_{n\to\infty} \log\tilde{K}_n - c(q_1(\Phi))\Gamma_n^{1/3}\log n = -\infty \tag{56}$$

Thus we have

$$\lim_{n\to\infty} \Pr(\exists j \in [\tilde{K}_n] : M_{r_j,j} \le \hat{\tau}_n) = 0 \tag{57}$$

Next, we look at $i \neq r_j$. Repeating the same steps above, we get

$$\beta_{i,j} \triangleq \Pr(|L_{i,j}(\Phi) - \Lambda_n q_0(\Phi)| \ge \varepsilon_n\Lambda_n) \tag{58}$$

$$\le 2^{-\Lambda_n D(q_0(\Phi)-\varepsilon_n \| q_0(\Phi))} + 2^{-\Lambda_n D(1-q_0(\Phi)-\varepsilon_n \| 1-q_0\Phi))} \tag{59}$$

$$= 2^{1-c(q_0(\Phi))\Gamma_n^{1/3}\log n + O(\varepsilon_n^3))} \tag{60}$$

Again, from the triangle inequality, we get

$$|L_{i,j}(\Phi) - \mu(\Phi)| \le 2\varepsilon_n = \hat{\tau}_n \tag{61}$$

From the union bound, we obtain

$$\Pr(\exists j \in [\tilde{K}_n]\, \exists i \in [n] \setminus \{r_j\} : |L_{i,j}(\Phi) - \mu(\Phi)| \ge \hat{\tau}_n) \tag{62}$$

$$\le \gamma + \sum_{j=1}^{\tilde{K}_n} \sum_{i \neq r_j} \beta_{i,j} \tag{63}$$

$$\le \gamma + n^2 2^{1-c(q_0(\Phi))\Gamma_n^{1/3}\log n + O(\varepsilon_n^3))} \tag{64}$$

Since $\Lambda_n = \omega(\log n)$, we have

$$\lim_{n\to\infty} \Pr(\exists j \in [\tilde{K}_n]\, \exists i \in [n] \setminus \{r_j\} : |L_{i,j}(\Phi) - \mu(\Phi)| \ge \hat{\tau}_n) = 0 \tag{65}$$

Thus, for any useful remapping Φ, the misdetection probability decays to zero as $n \to \infty$.

For any *useless* remapping Φ, following the same steps, one can prove that

$$\Pr(\text{Useless remapping } \Phi \text{ is inferred as useful.}) \tag{66}$$

$$\le \gamma + \sum_{i=1}^{n} \sum_{j=1}^{\tilde{K}_n} \Pr(M_{i,j} \ge \varepsilon_n\Lambda_n) \tag{67}$$

$$\le \gamma + n^2 2^{1-c(q_0(\Phi))\Gamma_n^{1/3}\log n + O(\varepsilon_n^3))} \tag{68}$$

$$= o_n(1) \tag{69}$$

Observing $|\mathfrak{S}(\mathfrak{X})| = |\mathfrak{X}|! = O_n(1)$ concludes the proof. □

C. De-Anonymization Scheme

In this section, we propose a de-anonymization scheme by combining the detection algorithms proposed in Section III-A-III-B, and performing a modified version of the typicality-based scheme proposed in [14]. Then using this scheme we prove the achievability of Theorem 1.

Given the database pair $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$ and the corresponding seed matrices $(\mathbf{G}^{(1)}, \mathbf{G}^{(2)})$, the de-anonymization scheme we propose is as follows:

1) Detect the replicas through Algorithm 1.

2) Remove all the extra replica columns from the seed matrix $\mathbf{G}^{(2)}$ to obtain $\tilde{\mathbf{G}}^{(2)}$ and perform seeded deletion detection via Algorithm 2 using $\mathbf{G}^{(1)}, \tilde{\mathbf{G}}^{(2)}$. At this step, we have an estimate $\hat{S}^n$ of the column repetition pattern S^n.

3) Based on $\hat{S}^n$ and the matching entries in $\mathbf{G}^{(1)}, \tilde{\mathbf{G}}^{(2)}$, obtain an estimate $\hat{p}_{X,Y^S|S}$ of $p_{X,Y^S|S}$ where

$$\hat{p}_X(x) \triangleq \frac{1}{\Lambda_n n} \sum_{i=1}^{\Lambda_n} \sum_{j=1}^{n} \mathbb{1}_{[G_{i,j}^{(1)}=x]}, \qquad \forall x \in \mathfrak{X} \tag{70}$$

$$\hat{p}_{Y|X}(y|x) = \frac{\sum_{i=1}^{\Lambda_n} \sum_{j=1}^{\tilde{K}_n} \mathbb{1}_{[G_{i,r_j}^{(1)}=x, \tilde{G}_{i,j}^{(2)}=y]}}{\sum_{i=1}^{\Lambda_n} \sum_{j=1}^{\tilde{K}_n} \mathbb{1}_{[\tilde{G}_{i,j}^{(2)}=y]}}, \quad \forall (x,y) \in \mathfrak{X}^2 \tag{71}$$

$$\hat{p}_S(s) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{[S_j=s]}, \qquad \forall s \geq 0 \tag{72}$$

and construct

$$\hat{p}_{X,Y^S|S}(x,y^s|s) = \begin{cases} \hat{p}_X(x) \mathbb{1}_{[y^s=*]} & \text{if } s=0 \\ \hat{p}_X(x) \prod_{j=1}^{s} \hat{p}_{Y|X}(y_j|x) & \text{if } s \geq 1 \end{cases} \tag{73}$$

with $y^s = y_1 \dots y_s$.

4) Using $\hat{S}^n$, place markers between the noisy replica runs of different columns to obtain $\tilde{\mathbf{D}}^{(2)}$. If a run has length 0, *i.e.* deleted, introduce a column consisting of erasure symbol $* \notin \mathfrak{X}$.

5) Fix $\varepsilon > 0$. Match the l^{th} row Y_l^K of $\tilde{\mathbf{D}}^{(2)}$ with the i^{th} row X_i^n of $\mathbf{D}^{(1)}$, if X_i is the only row of $\mathbf{D}^{(1)}$ jointly ε-typical [18, Chapter 7.6] with Y_l^K according to $\hat{p}_{X,Y^S,S}$, assigning $\hat{\Theta}_n(i) = l$. Otherwise, declare an error.

Let $\kappa_n^{(1)}$ and $\kappa_n^{(2)}$ be the error probabilities of the noisy replica detection (Algorithm 1) and the seeded deletion (Algorithm 2) algorithms, respectively. By the Law of Large Numbers, we have

$$\hat{p}_{X,Y^S|S} \xrightarrow{p} p_{X,Y^S|S} \tag{74}$$

and by the Continuous Mapping Theorem [25, Theorem 2.3] we have

$$\hat{H}(X,Y^S|S) \xrightarrow{p} H(X,Y^S|S) \tag{75}$$

$$I(\hat{X};\hat{Y}^{\hat{S}}|\hat{S}) \xrightarrow{p} I(X;Y^S|S) \tag{76}$$

where $\hat{H}(X,Y^S|S)$ and $\hat{I}(X,Y^S|S)$ denote the conditional joint entropy and conditional mutual information associated with $\hat{p}_{X,Y^S|S}$, respectively. Thus, for any $\varepsilon > 0$ we have

$$\kappa_n^{(3)} \triangleq \Pr(|\hat{H}(X,Y^S|S) - H(X,Y^S|S)| > \varepsilon) \xrightarrow{n\to\infty} 0 \tag{77}$$

$$\kappa_n^{(4)} \triangleq \Pr(|\hat{I}(X,Y^S|S) - I(X,Y^S|S)| > \varepsilon) \xrightarrow{n\to\infty} 0 \tag{78}$$

Now, from multiple triangle inequalities, the probability of error of the de-anonymization scheme can be bounded as

$$Pr(\text{error}) \leq 2^{-n(I(X;Y^S,S)-4\varepsilon-R)} + \varepsilon + \kappa_n^{(1)} + \kappa_n^{(2)} + \kappa_n^{(3)} + \kappa_n^{(4)} \tag{79}$$

$$\leq \varepsilon \tag{80}$$

as $n \to \infty$ as long as $R < I(X;Y^S,S) - 4\varepsilon$, concluding the proof of the main result.

IV. Conclusion

In this work, we have investigated the distribution-agnostic database de-anonymization problem under synchronization errors and noise. We have showed that the noisy replica detection algorithm of [14] tailored for specific $p_{X,Y}$ could be adjusted to work in tandem with a moment estimator to accommodate the unknown $p_{X,Y}$. We have proposed an outlier-detection-based seeded deletion detection algorithm and showed that a seed size growing double logarithmic with the number of rows is sufficient for correct estimation of the deletion pattern. Finally, we have used a joint-typicality-based de-anonymization scheme utilizing the estimated distributions. Overall, our results show that the resulting achievable database growth rate is equal to the matching capacity derived when full information on the underlying distributions is available.

References

[1] P. Ohm, "Broken promises of privacy: Responding to the surprising failure of anonymization," *UCLA L. Rev.*, vol. 57, p. 1701, 2009.

[2] F. M. Naini, J. Unnikrishnan, P. Thiran, and M. Vetterli, "Where you are is who you are: User identification by matching statistics," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 2, pp. 358–372, 2016.

[3] A. Datta, D. Sharma, and A. Sinha, "Provable de-anonymization of large datasets with sparse dimensions," in *International Conference on Principles of Security and Trust*. Springer, 2012, pp. 229–248.

[4] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. of IEEE Symposium on Security and Privacy*, 2008, pp. 111–125.

[5] L. Sweeney, "Weaving technology and policy together to maintain confidentiality," *The Journal of Law, Medicine & Ethics*, vol. 25, no. 2-3, pp. 98–110, 1997.

[6] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Matching anonymized and obfuscated time series to users' profiles," *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 724–741, 2019.

[7] D. Cullina, P. Mittal, and N. Kiyavash, "Fundamental limits of database alignment," in *Proc. of IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 651–655.

[8] F. Shirani, S. Garg, and E. Erkip, "A concentration of measure approach to database de-anonymization," in *Proc. of IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 2748–2752.

[9] O. E. Dai, D. Cullina, and N. Kiyavash, "Database alignment with gaussian features," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 3225–3233.

[10] D. Kunisky and J. Niles-Weed, "Strong recovery of geometric planted matchings," in *Proc. of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 2022, pp. 834–876.

[11] R. Tamir, "Joint Correlation Detection and Alignment of Gaussian Databases," *arXiv preprint arXiv:2211.01069*, 2022.

[12] S. Bakirtas and E. Erkip, "Database matching under column deletions," in *Proc. of IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 2720–2725.

[13] ——, "Matching of Markov databases under random column repetitions," in *2022 56th Asilomar Conference on Signals, Systems, and Computers*, 2022.

[14] ——, "Seeded database matching under noisy column repetitions," in *2022 IEEE Information Theory Workshop (ITW)*. IEEE, 2022, pp. 386–391.

[15] ——, "Database matching under noisy synchronization errors," *arXiv preprint arXiv:2301.06796*, 2023.

[16] ——, "Database matching under adversarial column deletions," in *2023 IEEE Information Theory Workshop (ITW)*. IEEE, 2023, pp. 181–185.

[17] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. MIT press, 2022.

[18] T. M. Cover, *Elements of Information Theory*. John Wiley & Sons, 2006.

[19] G. Morvai and B. Weiss, "Order estimation of Markov chains," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1496–1497, 2005.

[20] F. Shirani, S. Garg, and E. E., "Seeded graph matching: Efficient algorithms and theoretical guarantees," in *2017 51st Asilomar Conference on Signals, Systems, and Computers*, 2017, pp. 253–257.
[21] W. Blischke, "Moment estimators for the parameters of a mixture of two binomial distributions," *The Annals of Mathematical Statistics*, pp. 444–454, 1962.
[22] D. S. Moore and S. Kirkland, *The Basic Practice of Statistics*. WH Freeman New York, 2007, vol. 2.
[23] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004, vol. 26.
[24] R. B. Ash, *Information Theory*. Courier Corporation, 2012.
[25] A. W. Van der Vaart, *Asymptotic Statistics*. Cambridge University Press, 2000, vol. 3.