

Distribution-Agnostic Database De-Anonymization Under Obfuscation and Synchronization Errors

Serhat Bakirtas[✉], Member, IEEE, and Elza Erkip[✉], Fellow, IEEE

Abstract—Database de-anonymization typically involves matching an anonymized database with correlated publicly available data. Existing research focuses either on practical aspects without requiring knowledge of the data distribution yet provides limited guarantees, or on theoretical aspects assuming known distributions. This paper aims to bridge these two approaches, offering theoretical guarantees for database de-anonymization under synchronization errors and obfuscation without prior knowledge of data distribution. Using a modified replica detection algorithm and a new seeded deletion detection algorithm, we establish sufficient conditions on the database growth rate for successful matching, demonstrating a double-logarithmic seed size relative to row size is sufficient for detecting deletions in the database. Importantly, our findings indicate that these sufficient de-anonymization conditions are tight and are the same as in the distribution-aware setting, avoiding asymptotic performance loss due to unknown distributions. Finally, we evaluate the performance of our proposed algorithms through simulations, confirming their effectiveness in more practical, non-asymptotic, scenarios.

Index Terms—Dataset, database, matching, de-anonymization, alignment, distribution-agnostic, privacy, synchronization, obfuscation.

I. INTRODUCTION

THE accelerating growth of smart devices and applications has accelerated the collection of user-level micro-data both by private companies and public institutions. This data is often shared or sold after removing explicit user identifiers, a.k.a. *anonymization*, and coarsening of the data through noise, a.k.a. *obfuscation*. Despite these efforts, there is a growing concern about the privacy implications [1]. These concerns were further justified by the success of a series of practical de-anonymization attacks on real data [2], [3], [4], [5], [6]. In the light of these successful attacks, recently there has been an increasing effort on understanding the information-theoretic

Received 1 April 2024; revised 14 October 2024 and 3 February 2025; accepted 26 February 2025. Date of publication 10 March 2025; date of current version 26 March 2025. This work was supported in part by the National Science Foundation under Grant 2148293, Grant 2003182, and Grant 1815821; and in part by New York University (NYU) WIRELESS Industrial Affiliates. An earlier version of this paper was presented in part at the 2023 IEEE International Workshop on Information Forensics and Security (WIFS), Nuremberg, Germany [DOI: 10.1109/WIFS58808.2023.10374831]. The associate editor coordinating the review of this article and approving it for publication was Prof. Shouling Ji. (Corresponding author: Serhat Bakirtas.)

Serhat Bakirtas is with Qualcomm Inc., Boxborough, MA 01752 USA (e-mail: serhat.bakirtas@nyu.edu).

Elza Erkip is with the Department of Electrical and Computer Engineering, New York University, New York, NY 10003 USA (e-mail: elza@nyu.edu). Digital Object Identifier 10.1109/TIFS.2025.3550058

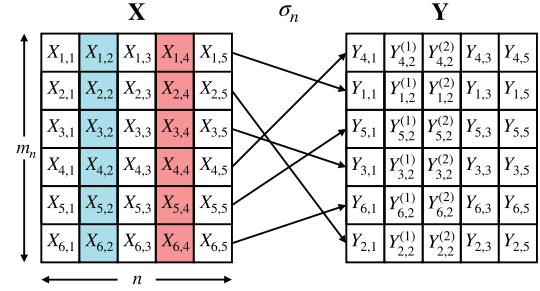


Fig. 1. An illustrative example of database matching under column repetitions. The column colored in red is deleted, whereas the column colored in blue is replicated. $Y_{i,2}^{(1)}$ and $Y_{i,2}^{(2)}$ denote noisy copies, *i.e.* replicas, of $X_{i,2}$. The goal of database de-anonymization studied in this paper is to estimate the correct row permutation $\sigma_n = (1 \ 2 \ 3 \ 4 \ 5 \ 6)$, by matching the rows of \mathbf{X} and \mathbf{Y} without *any prior information* on the underlying database (p_X), obfuscation ($p_{Y|X}$), and repetition (p_S) distributions.

and statistical foundations of *database de-anonymization*, a.k.a. *database alignment*, *database matching*, and *database recovery* [7], [8], [9], [10], [11], [12], [13], [14], [15], [16].

Our recent work focuses on the database de-anonymization problem under synchronization errors. In [13], we investigated the matching of Markov databases under synchronization errors only, with no subsequent obfuscation, *i.e.* noise, using a histogram-based detection method. In [14] and [15], we extended this effort to databases with noisy synchronization errors. We introduced a noisy replica detection algorithm, and a seeded deletion detection algorithm, and used joint-typicality-based matching to derive achievability results, which we subsequently showed to be tight, for a seed size double logarithmic with the row size of the database. In [13], [14], and [15] we assumed that the underlying distributions are known and tailored repetition detection and matching algorithms for these known distributions.

Motivated by most practical settings where the underlying distributions are not readily available, in this paper, we investigate the de-anonymization problem *without any prior knowledge of the underlying distributions*. Borrowing a noisy random column repetition model from [14], illustrated in Figure 1, we show that even though the distributions are not known *a priori*, matching can be performed with no asymptotic (in database size) loss of performance. To that end, we first modify the noisy replica detection algorithm proposed in [14] so that it still works in the 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), are available, the underlying deletion pattern could be inferred. Next, through a joint-typicality-based de-anonymization algorithm that relies on the estimated distributions and the repetition pattern, we derive a tight sufficient condition for database de-anonymization to succeed. Finally, we evaluate the performance of our proposed distribution-agnostic algorithms through simulations on finite databases and demonstrate their success in the non-asymptotic regime as well. We stress that, to the best of our knowledge, this is the first work that tackles the distribution-agnostic database matching under synchronization errors.

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. In Section IV we evaluate the performances of our proposed algorithms in the non-asymptotic regime via simulations. Section V consists of the concluding remarks. Proofs are provided in the Appendix.

Notation: We denote a matrix \mathbf{X} with bold capital letters, and its $(i, j)^{\text{th}}$ element with $X_{i,j}$. 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]. Unless stated otherwise, all logarithms are base 2. $H(\cdot)$ and $I(\cdot; \cdot)$ denote the Shannon entropy and the mutual information [18, Chapter 2], respectively. $D(p\|q)$ denotes the relative entropy [18, Chapter 2.3] (in bits) between two Bernoulli distributions with respective parameters p and q . \xrightarrow{p} denotes convergence in probability.

II. PROBLEM FORMULATION

In this section, we formalize our problem, adopting some definitions from [15]. In Section II-A, we formally define the database pairs of interest and seed matrices. In Section II-B, we present our key performance criterion, and state the main objective in this paper. In Section II-C, we discuss the set of assumptions used in our analyses. Finally, in Section II-D we present some well-known information-theoretic results that are used in this paper.

A. Anonymized and Labeled Correlated Databases & Seeds

Definition 1 (Anonymized Database): An (m_n, n, p_X) anonymized database $\mathbf{X} = \{X_{i,j} \in \mathcal{X}\}, (i, j) \in [m_n] \times [n]$ is a randomly generated $m_n \times n$ matrix with $X_{i,j} \stackrel{\text{i.i.d.}}{\sim} p_X$, where p_X has a finite discrete support $\mathcal{X} = \{1, \dots, |\mathcal{X}|\}$.

Definition 2 (Column Repetition Pattern): The column repetition pattern $S^n = \{S_1, S_2, \dots, S_n\}$ is a random vector with $S_i \stackrel{\text{i.i.d.}}{\sim} p_S$, where the repetition distribution p_S has a finite integer support $\{0, \dots, s_{\max}\}$. Here $\delta \triangleq p_S(0)$ is called the deletion probability.

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

Definition 4 (Obfuscation Distribution): The obfuscation (also referred to as *noise*) distribution $p_{Y|X}$ is a conditional probability distribution with both X and Y taking values from \mathcal{X} .



Fig. 2. Relation between the anonymized database \mathbf{X} and the labeled correlated database, \mathbf{Y} .

Definition 5 (Labeled Correlated Database): Let \mathbf{X} and S^n be a mutually-independent (m_n, n, p_X) anonymized database and repetition pattern, σ_n be an anonymization function, and $p_{Y|X}$ be an obfuscation distribution. Then, \mathbf{Y} is called the *labeled correlated database* if the i^{th} row X_i^n of \mathbf{X} and the $\sigma_n(i)^{\text{th}}$ row $Y_{\sigma_n(i)}^{K_n} = [Y_{\sigma_n(i),1}, \dots, Y_{\sigma_n(i),K_n}]$ of \mathbf{Y} have the relation given in (1)–(2), as shown at the bottom of the next page, where

$$K_j \triangleq \sum_{t=1}^j S_t. \quad (3)$$

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

The i^{th} row X_i of \mathbf{X} and the $\sigma_n(i)^{\text{th}}$ row $Y_{\sigma_n(i)}$ of \mathbf{Y} are called *matching rows*.

We can interpret (1) as follows: Since \mathbf{X} has *i.i.d.* entries and the noise on the each database entry is *i.i.d.*, each retained entry $X_{i,j}$ with $S_j \neq 0$ will correspond to an independent replica run $Y_{\sigma_n(i),K_{j-1}+1}, \dots, Y_{\sigma_n(i),K_j}$ in $Y_{\sigma_n(i)}$. Then, (2) directly follows from the fact that the noise on the retained entries is *i.i.d.*

The relationship between \mathbf{X} and \mathbf{Y} , as described in Definition 5, is illustrated in Figure 2.

As often done in both the graph matching [19] 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 6 (Seeds): Given a pair of anonymized and labeled correlated databases (\mathbf{X}, \mathbf{Y}) , a *seed* is a correctly matched row pair with the same underlying repetition pattern S^n . 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 and without loss of generality, we assume that the seed matrices $\mathbf{G}^{(1)}$ and $\mathbf{G}^{(2)}$ are not submatrices of \mathbf{X} and \mathbf{Y} . Throughout, 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 .

B. Performance Criterion

As the number of rows m_n increases for a fixed number of columns n , so does the probability of mismatch due to the increased number of candidates. In turn, in database de-anonymization problems, the relationship between the row size m_n and the column size n directly impacts the de-anonymization performance. Hence, as done in [8], [12], [13], [14], and [16], we utilize the database growth rate, defined below, as the main performance metric.

Definition 7 (Database Growth Rate): The *database growth rate* R of an (m_n, n, p_X) anonymized database is defined as

$$R = \lim_{n \rightarrow \infty} \frac{1}{n} \log m_n. \quad (4)$$

Definition 8 (Distribution-Agnostically Achievable Database Growth Rate): Consider a sequence of (m_n, n, p_X) anonymized databases \mathbf{X} with database growth rate R , an independent a repetition probability distribution p_S , an obfuscation distribution $p_{Y|X}$ and the resulting sequence of labeled correlated databases \mathbf{Y} . Given seed matrices $(\mathbf{G}^{(1)}, \mathbf{G}^{(2)})$ with a seed size Λ_n , the database growth rate R is said to be *distribution-agnostically achievable* if there exists a successful matching scheme $\psi : (\mathbf{X}, \mathbf{Y}, \mathbf{G}^{(1)}, \mathbf{G}^{(2)}) \mapsto \hat{\sigma}_n$ that does not rely on any knowledge on p_X , p_S and $p_{Y|X}$, with

$$\lim_{n \rightarrow \infty} \Pr(\sigma_n(I) \neq \hat{\sigma}_n(I)) = 0 \text{ where } I \sim \text{Unif}([m_n]) \quad (5)$$

where σ_n is true the anonymization function.

Remark 1: We note that the performance criterion described in Definition 8 is known as *almost-perfect recovery*. Another well-known performance criterion is the *perfect recovery* criterion where all rows need to be matched. More formally, the perfect recovery criterion corresponds to

$$\lim_{n \rightarrow \infty} \Pr(\sigma_n \neq \hat{\sigma}_n) = 0. \quad (6)$$

Observe that the difference between the two performance criteria is that in the almost-perfect criterion allows a sub-linear fraction of the database rows to be mismatched. This mismatch budget allows us to use tools such as typicality from information and communication theories. We refer to [10] for an in-depth analytical comparison of the Gaussian database matching results under these different performance criteria.

Definition 9 (Distribution-Agnostic Matching Capacity): The *distribution-agnostic matching capacity* C is the supremum of the set of all distribution-agnostically achievable database growth rates corresponding to a database distribution p_X , repetition distribution p_S , obfuscation distribution $p_{Y|X}$, and seed size Λ_n .

In this work, our main objective is the characterization of the *distribution-agnostic* matching capacity C . We remark that this is in contrast with prior work ([8], [12], [13], [14]) that assumed complete distributional knowledge.

Throughout, since we are interested in the distribution-agnostic matching capacity, we assume a positive database growth rate $R > 0$. In other words, we assume $n \sim \log m_n$.

C. Assumptions

Our analysis hinges on the following assumptions:

- (a) The anonymized database \mathbf{X} and the column repetition pattern S^n are known to be independent. It is also known that their components $X_{i,j}$ and S_i are *i.i.d.*, while the distributions p_X and p_S are not known.
- (b) The conditional independence of the noisy replicas stated in (2) is known, leading to a memoryless obfuscation model, whereas the noise distribution $p_{Y|X}$ is not.
- (c) $|\mathcal{X}|$ and s_{\max} are known.

As we argue in Sections III-A and III-B, S^n can be detected without any assumptions on the correlation between \mathbf{X} and S^n . In turn, one could test the independence of \mathbf{X} and the estimate \hat{S}^n of S^n . Furthermore, the *i.i.d.* nature of the components of \mathbf{X} and \hat{S}^n , and the obfuscation can be inferred via the Markov order estimation algorithm of [20] with a probability of error vanishing in n , justifying (a)-(b). Similarly, since $|\mathcal{X}|$ and s_{\max} do not depend on n , they can easily be estimated with a vanishing probability of error, justifying (c).

We note that our analysis is limited to the setting in which the anonymized database \mathbf{X} has *categorical* entries, namely they come from a discrete distribution, and synchronization errors and obfuscation occur *randomly* and not in an adversarial manner.

D. Preliminaries

In our derivations, we rely on some well-known classical information-theoretic definitions and results which we present below for the sake of completeness.

Definition 10 (Joint Entropy [18, Chapter 2.2]): The joint (Shannon) entropy associated with a pair of discrete random variables (X, Y) with a joint distribution $p_{X,Y}$ is defined as:

$$H(X, Y) \triangleq -\mathbb{E}[\log p_{X,Y}(X, Y)]. \quad (7)$$

Definition 11 (Joint Typicality [18, Chapter 7.6]): The ϵ -typical set $A_\epsilon^{(n)}(X, Y)$ associated with discrete random variables (X, Y) with joint distribution $p_{X,Y}$ is the set of all sequence pairs (x^n, y^n) satisfying

$$\left| -\frac{1}{n} \log p_{X^n, Y^n}(x^n, y^n) - H(X, Y) \right| \leq \epsilon \quad (8)$$

where $H(\mathcal{X}, \mathcal{Y})$ is the entropy rate of $(\mathcal{X}, \mathcal{Y})$.

Proposition 1 (Joint Asymptotic Equipartition Property [18, Theorem 7.6.1]): Let \tilde{X}^n and \tilde{Y}^n be generated according to the *i.i.d.* marginal distributions p_{X^n} and p_{Y^n} , independently. Then, the following holds:

$$\Pr((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}(X, Y)) \leq 2^{-n(I(X;Y)-3\epsilon)} \quad (9)$$

$$\Pr(Y_{\sigma_n(i)}^{K_n} = y^{K_n} | X_i^n = x^n) = \prod_{j: S_j \neq 0} \Pr(Y_{\sigma_n(i), K_{j-1}+1}, \dots, Y_{\sigma_n(i), K_j} = y_{K_{j-1}+1}, \dots, y_{K_j} | X_{i,j} = x_j) \quad (1)$$

$$= \prod_{j: S_j \neq 0} \prod_{s=1}^{S_j} p_{Y|X}(y_{K_{j-1}+s} | x_j) \quad (2)$$

where $I(X; Y) \triangleq H(X) + H(Y) - H(X, Y)$ is the mutual information.

III. MAIN RESULT

Our main result in Theorem 1 is on the distribution-agnostically achievable database growth rates when no prior information is provided on distributions p_X , $p_{Y|X}$, and p_S .

Theorem 1 (Distribution-Agnostic Matching Capacity): Consider an anonymized and labeled correlated database pair. We assume that the underlying database distribution p_X , the obfuscation function $p_{Y|X}$, and the column repetition distribution p_S are unknown. Given a seed size $\Lambda_n = \omega(\log n)$, the distribution-agnostic matching capacity is given by

$$C = I(X; Y^S | S) \quad (10)$$

where $S \sim p_S$, $X \sim p_X$ and $Y^S = Y_1, \dots, Y_S$ such that $Y_i | X \stackrel{\text{i.i.d.}}{\sim} p_{Y|X}$.

Theorem 1 implies 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 distributions p_X , $p_{Y|X}$ and p_S , and the actual column repetition pattern S^n a priori. Hence comparing Theorem 1 with [15, Theorem 1], we conclude that in the asymptotic regime where $n \rightarrow \infty$, not knowing the distributions and the realizations of the repetition pattern causes no loss in the matching capacity.

We stress that the proof of the converse part of Theorem 1 directly follows from [15, Theorem 1] since distribution-agnostic matching capacity cannot be better than the one that assumes distributional information. The rest of this section is on the proof of the achievability part of Theorem 1. In Section III-A, we present an algorithm for detecting noisy replicas and prove its asymptotic success. Then in Section III-B, we propose a seeded deletion detection algorithm and derive a sufficient seed size that guarantees its asymptotic success. Subsequently, in Section III-C, we present our de-anonymization algorithm that incorporates Sections III-A and III-B. Finally, in Section III-E, we focus on the special case of no obfuscation, where seeds are obsolete.

A. Distribution-Agnostic 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{Y} , denoted by W_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,

$$W_j \triangleq \sum_{i=1}^{m_n} \mathbb{1}_{[Y_{i,j+1} \neq Y_{i,j}]}, \quad \forall j \in [K_n - 1] \quad (11)$$

We first note that

$$W_j \sim \begin{cases} \text{Binom}(m_n, p_0), & \text{if } C_j^{(2)} \perp\!\!\!\perp C_{j+1}^{(2)} \\ \text{Binom}(m_n, p_1), & \text{otherwise} \end{cases} \quad (12)$$

where

$$p_0 \triangleq 1 - \sum_{y \in \mathcal{Y}} p_Y(y)^2 \quad (13)$$

Algorithm 1 Distribution-Agnostic Noisy Replica Detection Algorithm

Input : (\mathbf{Y}, m_n, K_n)
Output: isReplica

```

 $W \leftarrow \text{RunningHammingDist}(\mathbf{Y}); \quad /* \text{Eq. (11)} */$ 
 $(\hat{p}_0, \hat{p}_1) \leftarrow \text{EstimateParams}(W); \quad /* \text{Eq (66)-(67)} */$ 
 $\tau \leftarrow \frac{\hat{p}_0 + \hat{p}_1}{2}; \quad /* \text{Threshold} */$ 
isReplica  $\leftarrow \emptyset;$ 
/* Sweep over all  $(C_j^{(2)}, C_{j+1}^{(2)})$  pairs. */
for  $j = 1$  to  $K_n - 1$  do
    if  $W[j] \leq m_n \tau$  then
        /*  $C_j^{(2)}$  and  $C_{j+1}^{(2)}$  are correlated. */
        isReplica[j]  $\leftarrow$  TRUE;
    else
        /*  $C_j^{(2)}$  and  $C_{j+1}^{(2)}$  are independent. */
        isReplica[j]  $\leftarrow$  FALSE;
    end
end

```

$$p_1 \triangleq 1 - \sum_{x \in \mathcal{X}} p_X(x) \sum_{y \in \mathcal{Y}} p_{Y|X}(y|x)^2. \quad (14)$$

Here, p_0 is the probability that two independent entries of \mathbf{X} have different noisy observations in \mathbf{Y} , and p_1 is the probability that the same replicated entry of \mathbf{X} has two different noisy observations in \mathbf{Y} .

From [16, Lemma 1], we know that as long as the databases are dependent, *i.e.*, $p_{X,Y} \neq p_X p_Y$, we have $p_0 > p_1$ for any $p_{X,Y}$ suggesting that replicas can be potentially detected based on W_j similar to [14] and [16]. However, the algorithm in [14] relies on a threshold test with the threshold depending on $p_{X,Y}$ through p_0 and p_1 . In Algorithm 1, we propose the following modification for the distribution-agnostic setting: 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 [21] and then use the algorithm proposed in [14, Section III-A] for replica detection. Note that we can use Blischke's estimator because the Binomial mixture is guaranteed to have two distinct components. More formally, the distribution of W_j conditioned on S^n is given by

$$\Pr(W_j = w | S^n) = \binom{m_n}{w} [\alpha p_0^w (1 - p_0)^{m_n - w} + (1 - \alpha) p_1^w (1 - p_1)^{m_n - w}] \quad (15)$$

for $w = 0, \dots, m_n$ where the mixing parameter α is given by

$$\alpha = \frac{1}{K_n - 1} \left(n - \sum_{j=1}^n \mathbb{1}_{[S_j=0]} \right). \quad (16)$$

Since p_S , and in turn α are constant in n , it can easily be verified that as $n \rightarrow \infty$, $\alpha \xrightarrow{P} \frac{1-\delta}{\mathbb{E}[S]}$. Hence α is bounded away from both 0 and 1, suggesting that the moment estimator of [21] and in turn, Algorithm 1 can be used to detect the replicas.

Lemma 1 (Noisy Replica Detection): Let E_j denote the event that Algorithm 1 fails to infer the correct relationship

between $C_j^{(2)}$ and $C_{j+1}^{(2)}$, $j = 1, \dots, K_n - 1$. Then, given $m_n = \omega(\log n)$

$$\kappa_n^{(1)} \triangleq \Pr\left(\bigcup_{j=1}^{K-1} E_j\right) \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (17)$$

Proof: See Appendix A. \square

Note that the condition in Lemma 1 is automatically satisfied since m_n is exponential in n (Definition 7). Furthermore, Algorithm 1 has a runtime of $\mathcal{O}(m_n n)$, the computational bottleneck being the computation of the running Hamming distances W_j . Finally, we stress that as opposed to deletion detection, discussed in Section III-B, no seeds are necessary for replica detection.

B. Distribution-Agnostic Deletion Detection Using Seeds

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

$$L_{i,j} \triangleq \sum_{t=1}^{\Lambda_n} \mathbb{1}_{[G_{i,t}^{(1)} \neq G_{i,t}^{(2)}]} \quad (18)$$

Observe that

$$L_{i,j} \sim \begin{cases} \text{Binom}(\Lambda_n, q_0), & G_i^{(1)} \perp\!\!\!\perp G_j^{(2)} \\ \text{Binom}(\Lambda_n, q_1), & \text{otherwise} \end{cases} \quad (19)$$

where

$$q_0 = 1 - \sum_{x \in \mathcal{X}} p_X(x)p_Y(x) \quad (20)$$

$$q_1 = 1 - \sum_{x \in \mathcal{X}} p_{X,Y}(x, x). \quad (21)$$

Here, q_0 is the probability that two independent (non-matching) entries in $\mathbf{G}^{(1)}$ and $\mathbf{G}^{(2)}$ are different, and q_1 is the probability that two matching entries in $\mathbf{G}^{(1)}$ and $\mathbf{G}^{(2)}$ are different.

Thus, we have a problem seemingly similar to the one in Section III-A. However, we cannot utilize similar tools here because of the following:

- (i) Recall that the two components p_0 and p_1 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 dependent, *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 $X_{i,j} \sim \text{Unif}(\mathcal{X})$, and the transition matrix \mathbf{P} associated with $p_{Y|X}$ has unit trace. Then,

$$q_0 - q_1 = \sum_{x \in \mathcal{X}} p_{X,Y}(x, x) - p_X(x)p_Y(x) \quad (22)$$

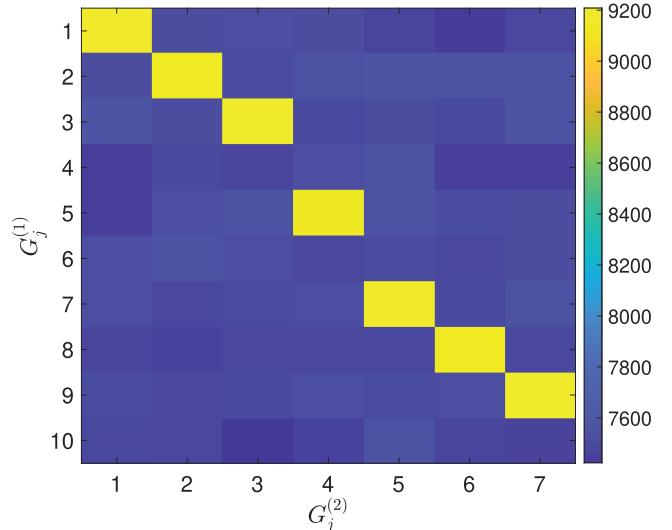


Fig. 3. Hamming distances between the columns of $\mathbf{G}^{(1)}$ and $\mathbf{G}^{(2)}$ with $n = 10$, $\tilde{K}_n = 7$ and $\Lambda_n = 10^4$ for $q_0 \approx 0.76$ and $q_1 \approx 0.92$. The $(i, j)^{\text{th}}$ element corresponds to $L_{i,j}$, with the color bar indicating the approximate values. It can be seen that there are no outliers in the 4th, 6th, and 10th rows. Hence, it can be inferred that $I_{\text{del}} = (4, 6, 10)$.

$$= \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} p_{Y|X}(x|x) - p_Y(x) \quad (23)$$

$$= \frac{1}{|\mathcal{X}|} (\text{tr}(\mathbf{P}) - 1) \quad (24)$$

$$= 0 \quad (25)$$

In [14], this problem was solved using the following modification: Based on $p_{X,Y}$, a bijective remapping $\Phi \in \mathcal{S}(\mathcal{X})$ is picked and applied to all the entries of $\mathbf{G}^{(2)}$ to obtain $\mathbf{G}^{(2)}(\Phi)$ before computing the Hamming distances $L_{i,j}$, where $\mathcal{S}(\mathcal{X})$ denotes the symmetry group of \mathcal{X} . Denoting the resulting version of the Hamming distance by $L_{i,j}(\Phi)$, where

$$L_{i,j}(\Phi) \triangleq \sum_{t=1}^{\Lambda_n} \mathbb{1}_{[G_{i,t}^{(1)} \neq G_{i,t}^{(2)}(\Phi)]} \quad (26)$$

it was proved in [14, Lemma 2] that there as long as $p_{X,Y} \neq p_X p_Y$, there exists $\Phi \in \mathcal{S}(\mathcal{X})$ such that $q_0(\Phi) \neq q_1(\Phi)$. We will call such Φ a *useful remapping*.

- (ii) In the known distribution setting, we chose a 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^{th} row, whereas the remaining $n - 1$ rows will contain $\text{Binom}(\Lambda_n, q_0(\Phi))$ components, as illustrated in Figure 3. Hence, it can be seen that the mixture parameter β of this Binomial mixture distribution approaches 1 since

$$\beta = \frac{(n-1)\tilde{K}_n}{n\tilde{K}_n} = 1 - \frac{1}{n}. \quad (27)$$

Algorithm 2 Distribution-Agnostic Seeded Deletion Detection Algorithm

Input : $(\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \Lambda_n, n, \tilde{K}_n, \mathcal{X})$

Output: \hat{I}_R ; /* Retention set */

$\mathcal{S}(\mathcal{X}) \leftarrow \text{SymmetryGroup}(\mathcal{X})$;

$\hat{\tau}_n \leftarrow 2\Lambda_n^{2/3}(\log n)^{1/3}$; /* Threshold */

/* Sweep over Φ until a useful remapping is found. */

for $s \leftarrow 1$ **to** $|\mathcal{X}|$ **do**

$\hat{I}_R \leftarrow \emptyset$;

$\Phi \leftarrow \mathcal{S}(\mathcal{X})[s]$; /* Pick a remapping. */

$\mathbf{L}(\Phi) \leftarrow \text{HammDist}(\mathbf{G}^{(1)}, \mathbf{G}^{(2)}(\Phi))$; /* Eq. (18) */

$\mu(\Phi) \leftarrow \text{SampleMean}(\mathbf{L}(\Phi))$; /* Eq. (28) */

$\mathbf{M}(\Phi) \leftarrow |\mathbf{L}(\Phi) - \mu(\Phi)|$; /* Eq. (29) */

/* Sweep over $G_j^{(2)}(\Phi)$. */

for $j \leftarrow 1$ **to** \tilde{K}_n **do**

$\text{count} \leftarrow 0$;

/* Sweep over $G_i^{(1)}$. */

for $i \leftarrow 1$ **to** n **do**

if $\mathbf{M}(\Phi)[i][j] \leq \hat{\tau}_n$ **then**

/* $G_i^{(1)}$ and $G_j^{(2)}(\Phi)$ are correlated. */

/* Add $G_i^{(1)}$ to the retention set. */

$\hat{I}_R \leftarrow \hat{I}_R \cup \{i\}$;

$\text{count} \leftarrow \text{count} + 1$;

end

end

if $\text{count} > 1$ **then**

/* $G_i^{(1)}$ matched with multiple $G_j^{(2)}(\Phi)$ */

return ERROR; /* Misdetection error. */

else

if $\text{count} = 0$ **then**

/* No outliers found (Φ is useless). */

 Skip to next Φ ;

end

end

end

return \hat{I}_R ;

end

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

In Algorithm 2, we 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 $G_j^{(2)}(\Phi)$ being dependent, and in turn $S_j \neq 0$. On the other hand, the lack of outliers in any given column of $\mathbf{L}(\Phi)$ implies that Φ is not useful. Thus, Algorithm 2 is capable of deciding whether a given remapping is useful or not. In fact, the algorithm sweeps over all elements of $\mathcal{S}(\mathcal{X})$ until we encounter a useful one.

To detect the outliers in $\mathbf{L}(\Phi)$, we propose to use the absolute deviations $\mathbf{M}(\Phi)$, *i.e.*, distances of $L_{i,j}(\Phi)$ to the

sample mean $\mu(\Phi)$ of $\mathbf{L}(\Phi)$. More formally, we have

$$\mu(\Phi) \triangleq \frac{1}{n\tilde{K}_n} \sum_{i=1}^n \sum_{j=1}^{\tilde{K}_n} L_{i,j}(\Phi) \quad (28)$$

$$M_{i,j}(\Phi) \triangleq |L_{i,j}(\Phi) - \mu(\Phi)|, \quad \forall (i, j) \in [n] \times [\tilde{K}_n] \quad (29)$$

In Algorithm 2, we test these absolute deviations $M_{i,j}(\Phi)$ against a threshold $\hat{\tau}_n$ independent of the underlying distribution $p_{X,Y}$. If $M_{i,j}(\Phi)$ is 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 absolute deviations are also divided by the sample standard deviation. In Algorithm 2, for the sake of notational brevity, we will avoid such division.

We observe that the runtime of Algorithm 2 is $\mathcal{O}(\Lambda_n n^2)$ due to the computation of $\mathbf{L}(\Phi)$.

Lemma 2 below 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 set of indices of retained columns and \hat{I}_R be its estimate output by Algorithm 2. Then for any seed size $\Lambda_n = \omega(\log n)$, we have*

$$\kappa_n^{(2)} \triangleq \lim_{n \rightarrow \infty} \Pr(\hat{I}_R \neq I_R) = 0. \quad (30)$$

Proof: See Appendix B. □

C. Distribution-Agnostic De-Anonymization Scheme

In this section, we propose a de-anonymization scheme by combining Algorithm 1 and Algorithm 2, and performing a modified version of the typicality-based de-anonymization proposed in [14]. This then leads to the achievability proof of Theorem 1 in Section III-D.

Given the database pair (\mathbf{X}, \mathbf{Y}) and the corresponding seed matrices $(\mathbf{G}^{(1)}, \mathbf{G}^{(2)})$, the proposed de-anonymization scheme given in Algorithm 3 is as follows:

- (i) Detect the replicas through Algorithm 1.
- (ii) 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 .
- (iii) Based on \hat{S}^n and the matching entries in $\mathbf{G}^{(1)}, \tilde{\mathbf{G}}^{(2)}$, obtain the maximum likelihood estimate [23] $\hat{p}_{X,Y|S}$ of $p_{X,Y|S}$ where

$$\hat{p}_X(x) \triangleq \frac{1}{\Lambda_n n} \sum_{i=1}^{\Lambda_n} \sum_{j=1}^n \mathbb{1}_{[G_{i,j}^{(1)}=x]}, \quad \forall x \in \mathcal{X} \quad (31)$$

$$\hat{p}_{Y|X}(y|x) \triangleq \frac{\sum_{i=1}^{\Lambda_n} \sum_{j=1}^{\tilde{K}_n} \mathbb{1}_{[G_{i,j}^{(1)}=x, \tilde{G}_{i,j}^{(2)}=y]}}{\sum_{i=1}^{\Lambda_n} \sum_{j=1}^{\tilde{K}_n} \mathbb{1}_{[\tilde{G}_{i,j}^{(2)}=y]}}, \quad \forall (x, y) \in \mathcal{X}^2 \quad (32)$$

$$\hat{p}_S(s) \triangleq \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{[S_j=s]}, \quad \forall s \geq 0 \quad (33)$$

Algorithm 3 Distribution-Agnostic Database De-Anonymization Scheme

Input : $(\mathbf{X}, \mathbf{Y}, \epsilon, \mathbf{G}^{(1)}, \mathbf{G}^{(2)})$
Output: $\hat{\sigma}_n$; /* Anonymization function */
/* Noisy Replica Detection (Sec. II-A) */
 $\text{isReplica} \leftarrow \text{Alg.1}(\mathbf{Y})$; /* (i). */
/* Seeded Deletion Detection (Sec. II-B) */
 $\text{isDeleted} \leftarrow \text{Alg.2}(\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \text{isReplica})$; /* (ii). */
 $\hat{S}^n \leftarrow \text{EstimateRepetitionPattern}(\text{isReplica}, \text{isDeleted})$;
/* Estimate the underlying distribution. */
 $\hat{p}_{X, Y^S|S} \leftarrow \text{EstimateDist}(\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \hat{S}^n)$; /* (iii). */
 $\tilde{\mathbf{Y}} \leftarrow \text{MarkerAddition}(\mathbf{Y}, \hat{S}^n)$; /* (iv). */
/* Sweep over all row pairs (X_i, Y_j) . */
for $i = 1$ **to** $\text{rowSize}(\mathbf{X})$ **do**
 count $\leftarrow 0$;
 for $j = 1$ **to** $\text{rowSize}(\tilde{\mathbf{Y}})$ **do**
 if $\text{isJointTypical}(\mathbf{X}[i][:], \mathbf{Y}[j][:], \hat{S}^n, \hat{p}_{X, Y^S|S}, \epsilon)$
 then
 /* X_i and Y_j are matching rows. */
 $\hat{\sigma}_n[i] \leftarrow j$;
 count $\leftarrow \text{count} + 1$;
 end
 end
 if $\text{count} \neq 1$ **then**
 /* X_i is matched with multiple Y_j . */
 $\hat{\sigma}_n[i] \leftarrow 0$; /* Matching error. */
 end
 end

and construct

$$\hat{p}_{X, Y^S|S}(x, y^s|s) = \begin{cases} \hat{p}_X(x) \mathbb{1}_{[y^s = *]} & \text{if } s = 0 \\ \hat{p}_X(x) \prod_{j=1}^s \hat{p}_{Y|X}(y_j|x) & \text{if } s \geq 1 \end{cases} \quad (34)$$

with $y^s = y_1 \dots y_s$ and $*$ denoting erasure.

- (iv) Using \hat{S}^n , place markers between the noisy replica runs of different columns to obtain $\tilde{\mathbf{Y}}$. If a run has length 0, i.e. deleted, introduce a column consisting of erasure symbol $*$ $\notin \mathcal{X}$.
(v) Fix $\epsilon > 0$. Match the i^{th} row $Y_l^{K_n}$ of $\tilde{\mathbf{Y}}$ with the i^{th} row X_i^n of \mathbf{X} , if X_i is the only row of \mathbf{X} jointly ϵ -typical (See Definition 11) with $Y_l^{K_n}$ according to $\hat{p}_{X, Y^S|S}$, assigning $\hat{\sigma}_n(i) = l$. Otherwise, declare an error.

Finally, note that the runtime of Algorithm 3 is $\mathcal{O}(m_n^2 n)$ due to the typicality check (each $\mathcal{O}(n)$) for all row pairs $(X_i^n, Y_l^{K_n})$, $(i, j) \in [m_n]^2$.

D. Proof of Theorem 1

We are now ready to prove Theorem 1 where we use Algorithm 3 to obtain a matching scheme $\hat{\sigma}_n$.

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. Using (31)–(33) and the Law of Large Numbers, we have

$$\hat{p}_{X, Y^S|S} \xrightarrow{p} p_{X, Y^S|S} \quad (35)$$

and by the Continuous Mapping Theorem [24, Theorem 2.3] we have

$$\hat{H}(X, Y^S|S) \xrightarrow{p} H(X, Y^S|S) \quad (36)$$

$$I(\hat{X}; \hat{Y}^S|\hat{S}) \xrightarrow{p} I(X; Y^S|S) \quad (37)$$

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

$$\kappa_n^{(3)} \triangleq \Pr(|\hat{H}(X, Y^S|S) - H(X, Y^S|S)| > \epsilon) \xrightarrow{n \rightarrow \infty} 0 \quad (38)$$

$$\kappa_n^{(4)} \triangleq \Pr(|\hat{I}(X, Y^S|S) - I(X, Y^S|S)| > \epsilon) \xrightarrow{n \rightarrow \infty} 0 \quad (39)$$

Using Proposition 1 and a series of union bounds and triangle inequalities involving $\hat{H}(X, Y^S|S)$, $H(X, Y^S|S)$, $\hat{I}(X, Y^S|S)$, and $I(X, Y^S|S)$, the probability of error of the de-anonymization scheme (See (5).) can be bounded as

$$\Pr(\text{Error}) \leq 2^{-n(I(X; Y^S|S) - 4\epsilon - R)} + \epsilon + \sum_{i=1}^4 \kappa_n^{(i)} \quad (40)$$

$$\leq 2\epsilon \quad (41)$$

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

E. No-Obfuscation Setting

In Sections III-A and III-B, we focused on the obfuscated (noisy) setting and proposed algorithms for detecting noisy replicas (Algorithm 1) and column deletions (Algorithm 2). As discussed in Sections III-A and III-B, the key idea behind detection is either extracting permutation-invariant features of the columns (Algorithm 1) or assuming the correct matching is given for the seeds (Algorithm 2). We showed that in the general noisy setting, for the latter approach to succeed a double-logarithmic seed size Λ_n is sufficient.

Now, we will focus on the no-obfuscation (repetition-only) setting, where

$$p_{Y|X}(y|x) = \mathbb{1}_{[y=x]}, \quad \forall (x, y) \in \mathcal{X}^2 \quad (42)$$

as a special case. Following a similar approach to Algorithm 1, through the extraction of a new set of permutation-invariant features of the database columns, we argue that seeds are not required for repetition (both *replica* and *deletion*) detection. Specifically, we will replace Algorithm 1 and Algorithm 2 a single the histogram-based detection algorithm of [16] which works as follows:

- (i) First, we construct the respective column histogram matrices $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$ and of \mathbf{X} and \mathbf{Y} as follows:

$$H_{i,j}^{(1)} = \sum_{t=1}^{m_n} \mathbb{1}_{[X_{t,j}=i]}, \quad \forall i \in \mathcal{X}, \forall j \in [n] \quad (43)$$

$$H_{i,j}^{(2)} = \sum_{t=1}^{m_n} \mathbb{1}_{[Y_{t,j}=i]}, \quad \forall i \in \mathcal{X}, \forall j \in [K_n] \quad (44)$$

- (ii) Next, we count the number of times the j^{th} column $H_j^{(1)}$ of $\mathbf{H}^{(1)}$ is present in $\mathbf{M}^{(2)}$. If $H_j^{(1)}$ is present s times in $\mathbf{H}^{(2)}$, we assign $\hat{S}_j = s$, where \hat{S}^n is the estimate of the

repetition pattern S^n . If $H_j^{(1)}$ is absent in $\mathbf{H}^{(2)}$, we assign $\hat{S}_j = 0$.

Note that a repetition error could occur only when there are identical columns in $\mathbf{H}^{(1)}$ whose probability, we argue, vanishes with the column size n in the following lemma:

Lemma 3 (Asymptotic Uniqueness of the Column Histograms [16, Lemma 1]): Let $H_j^{(1)}$ denote the histogram of the j^{th} column of \mathbf{X} . Then, for any $m_n = \omega\left(n^{\frac{4}{|\mathcal{X}|-1}}\right)$

$$\lim_{n \rightarrow \infty} \Pr\left(\exists i, j \in [n], i \neq j, H_i^{(1)} = H_j^{(1)}\right) \rightarrow 0. \quad (45)$$

Note that since m_n is exponential in n (See Definition 7), $m_n = \omega(n^s) \forall s \in \mathbb{Q}$. Hence, the order relation given in Lemma 3 is automatically satisfied and hence repetition detection could be performed without any seeds in the no-obfuscation setting with a runtime of $\mathcal{O}(m_n n)$ due to the computation of the histogram matrices $(\mathbf{H}^{(1)}, \mathbf{H}^{(2)})$.

For the non-obfuscation setting, we also consider a modified de-anonymization scheme, given in Algorithm 4, that does not rely on any estimates of the underlying distributions. Specifically, we perform de-anonymization via exact sequence matching as follows:

- First, we perform repetition detection via the histogram-based detection algorithm described above.
- Next, we discard the deleted columns from \mathbf{X} to obtain $\bar{\mathbf{X}}$.
- Similarly, we discard all additional copies in a repetition run in \mathbf{Y} to obtain $\bar{\mathbf{Y}}$.
- Finally, we match the i^{th} row X_i^n with the j^{th} row $Y_j^{K_n}$, assigning $\hat{\sigma}_n(i) = j$, if $\bar{Y}_j^{K_n}$ is the only row of $\bar{\mathbf{Y}}$ equal to $\bar{X}_i^{K_n}$.

The histogram-based detection and the modified de-anonymization scheme are given in Algorithm 4. We note that the overall runtime of Algorithm 4 is $\mathcal{O}(m_n^2 n)$ due to the exact row matching (each $\mathcal{O}(n)$) of m_n^2 row pairs.

The matching scheme described above and given in Algorithm 4 leads to the achievability result given in Theorem 2. The converse immediately follows from the distribution-aware matching capacity result of [15, Theorem 2].

Theorem 2 (Distribution-Agnostic Matching Capacity in the No-Obfuscation Setting): Consider an anonymized and labeled correlated database pair. We assume that the underlying database distribution p_X and the column repetition distribution p_S are unknown. Suppose there is no obfuscation, i.e.,

$$p_{Y|X}(y|x) = \mathbb{1}_{[y=x]}, \quad \forall (x, y) \in \mathcal{X}^2. \quad (46)$$

Then, for any seed size Λ_n , the distribution-agnostic matching capacity C is given by

$$C = (1 - \delta)H(X). \quad (47)$$

Proof: See Appendix C. \square

Observe that Theorem 2 states that the repetition distribution p_S appears in the matching capacity only through the deletion probability δ in the no-obfuscation setting.

Algorithm 4 Distribution-Agnostic De-Anonymization In No-Obfuscation Setting

```

Input :  $(\mathbf{X}, \mathbf{Y})$ 
Output:  $\hat{\sigma}_n$ ; /* Anonymization function */
/* Compute column histograms of  $\mathbf{H}^{(1)}$  and  $\mathbf{H}^{(2)}$ . */
 $(\mathbf{H}^{(1)}, \mathbf{H}^{(2)}) \leftarrow \text{ColHist}(\mathbf{X}, \mathbf{Y})$ ; /* Eq. (43)-(44). */
/* Histogram-based repetition detection */
/* Sweep over all  $H_i^{(1)}$ . */
for  $i = 1$  to  $\text{columnSize}(\mathbf{H}^{(1)})$  do
  /* Count how many times  $H_i^{(1)}$  is in  $\mathbf{H}^{(2)}$ . */
   $\text{count} \leftarrow 0$ ;
  for  $j = 1$  to  $\text{columnSize}(\mathbf{H}^{(2)})$  do
    if  $\mathbf{H}^{(2)}[:, j] = \mathbf{H}^{(1)}[:, i]$  then
       $\text{count} \leftarrow \text{count} + 1$ ;
    end
  end
   $\hat{S}[i] \leftarrow \text{count}$ ;
end
 $(\bar{\mathbf{X}}, \bar{\mathbf{Y}}) \leftarrow \text{ColumnDiscard}(\mathbf{X}, \mathbf{Y}, \hat{S})$ 
/* Exact sequence matching */
/* Sweep over all row pairs  $(X_i, Y_j)$ . */
for  $i = 1$  to  $\text{rowSize}(\mathbf{X})$  do
   $\text{count} \leftarrow 0$ ;
  for  $j = 1$  to  $\text{rowSize}(\mathbf{Y})$  do
    if  $\bar{\mathbf{X}}[i][:] = \bar{\mathbf{Y}}[j][:]$  then
      /*  $X_i$  and  $Y_j$  are matching rows. */
       $\hat{\sigma}_n[i] \leftarrow j$ ;
       $\text{count} \leftarrow \text{count} + 1$ ;
    end
  end
  if  $\text{count} \neq 1$  then
    /*  $X_i$  is matched with multiple  $Y_j$ . */
     $\hat{\sigma}_n[i] \leftarrow 0$ ; /* Matching error. */
  end
end

```

IV. NON-ASYMPTOTIC REGIME

In Section III, we devised various detection and de-anonymization algorithms and proved their performances (in terms of error probabilities and matching capacity) in the asymptotic regime where the column size n grows to infinity. In this section, through extensive experiments, we will study the performances of these algorithms when n is finite.

Throughout this section, unless stated otherwise, we will focus on the following simulation configuration:

- *Database size* (m_n, n): We choose column size $n = 100$ with the exception of the evaluation of the de-anonymization scheme in Section IV-C where we take $n = 25$. We evaluate the performance of our algorithms for a range of values of row sizes m_n .
- *Database distribution* p_X : Without loss of generality, we focus on the uniform distribution $p_X(x) = \frac{1}{|\mathcal{X}|}, \forall x \in \mathcal{X}$ where we arbitrarily take the alphabet size $|\mathcal{X}| = 5$. In Section IV-D, we will sweep over $|\mathcal{X}|$.
- *Obfuscation* $p_{Y|X}$: We will consider the $|\mathcal{X}|$ -ary symmetric channel with a crossover probability ϵ . More formally, for

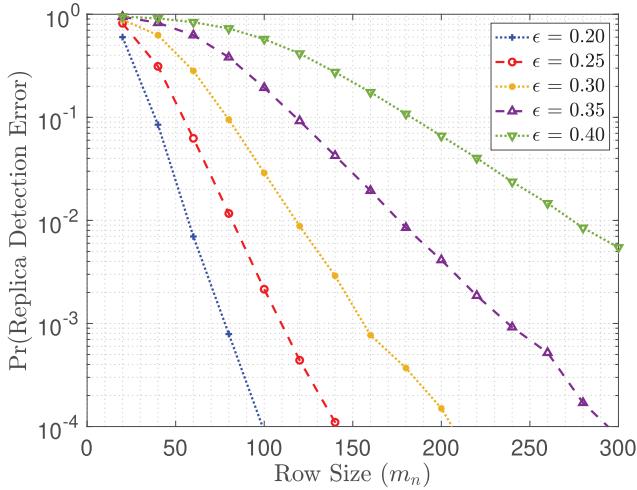


Fig. 4. Probability of error of the noisy replica detection algorithm (Algorithm 1) $\kappa_n^{(1)}$ vs. the row size m_n with 10^5 trials. The y-axis is given in logarithmic scale to validate the exponential relation between the error probability and m_n given in (72). Different curves correspond to different crossover probabilities ϵ .

any $(x, y) \in \mathcal{X}^2$, we have

$$p_{Y|X}(y|x) = \begin{cases} 1 - \epsilon, & \text{if } y = x \\ \frac{\epsilon}{|\mathcal{X}| - 1}, & \text{if } y \neq x \end{cases} \quad (48)$$

where we simulate with a range of values of ϵ .

- *Repetition distribution p_S :* We consider a “deletion-duplication” model for the repetition distribution with

$$p_S(s) = \begin{cases} \delta, & \text{if } s = 0 \\ 1 - \delta - \gamma, & \text{if } s = 1 \\ \gamma, & \text{if } s = 2 \end{cases} \quad (49)$$

where $\delta = 0.3$ and $\gamma = 0.2$.

We stress that even though we assume a discrete uniform distribution for p_X , and a symmetric noise model for $p_{Y|X}$ for our experimental analyses, as we have proved in Section III, our algorithms work for *any* discrete distribution p_X with finite support \mathcal{X} , and *any* obfuscation distribution $p_{Y|X}$ where X and Y take values from \mathcal{X} .

A. Noisy Replica Detection

Figure 4 demonstrates the relationship between the replica detection error $\kappa_n^{(1)}$ of Algorithm 1 and the row size m_n . Note that the linear relation in the semi-logarithmic plot validates the exponential decay of $\kappa_n^{(1)}$ with m_n stated in (72).

Figure 4 shows that as the crossover probability ϵ increases, *i.e.*, the labeled correlated database \mathbf{Y} gets noisier, the error probability decay becomes slower, leading to a worse detection performance, as expected. This is due to the following: Our simulation configuration leads to the following Binomial parameters p_0 (Eq. (13)) and p_1 (Eq. (14)):

$$p_0 = 1 - \frac{1}{|\mathcal{X}|} \quad (50)$$

$$p_1 = 1 - (1 - \epsilon)^2 - \frac{\epsilon^2}{|\mathcal{X}| - 1} \quad (51)$$

Consequently,

$$p_0 - p_1 = (1 - \epsilon)^2 + \frac{\epsilon^2}{|\mathcal{X}| - 1} - \frac{1}{|\mathcal{X}|}, \quad (52)$$

and $D(\tau||p_0)$ and $D(1 - \tau||1 - p_1)$ are decreasing functions of ϵ for any $\epsilon < 1 - \frac{1}{|\mathcal{X}|}$. Thus, we have a worse replica detection performance for a higher crossover probability ϵ .

B. Seeded Deletion Detection

While our replica detection algorithm (Algorithm 1) works for a small column size as shown in Figure 4, our seeded deletion detection algorithm (Algorithm 2) requires large column sizes. This is because the threshold $\hat{\tau}_n$ of Algorithm 2 is chosen based on its asymptotic properties and does not take estimated parameters into account, unlike the threshold $m_n\tau$ of Algorithm 1. Hence, while the outliers still exist in $\mathbf{L}(\Phi)$ (Eq. (26)) for any useful remapping Φ , the threshold becomes too large to distinguish the outlier from the rest of the entries.

To overcome this problem, we propose a modification to Algorithm 2 based on the following observation: As shown for a family of distributions in [25], the largest two order statistics of a sequence of n *i.i.d.* random variables typically have close values for large n . More formally, let $T_{(1),j} \leq \dots \leq T_{(n),j}$ be the order statistics of the j^{th} column of $\mathbf{M}(\Phi)$ (Eq. (29)) and define

$$R_{(i),j} \triangleq \frac{T_{(i),j}}{T_{(i-1),j}}, \quad 2 \leq i \leq n. \quad (53)$$

Then, for large n , we expect to have $T_{(n),j} = M_{r_j,j}(\Phi)$, where r_j is the index of the j^{th} retained column (Eq. (76)), and

$$\lim_{n \rightarrow \infty} R_{(n),j} = \infty \quad (54)$$

$$\lim_{n \rightarrow \infty} R_{(n-1),j} = 1 \quad (55)$$

Based on this, we propose the following modification: First, we compute the sample means $\bar{R}_{(n)}$ and $\bar{R}_{(n-1)}$ of $\{R_{(n),j}\}_{j=1}^{\hat{K}_n}$ and $\{R_{(n-1),j}\}_{j=1}^{\hat{K}_n}$, respectively. Then, the algorithm declares Φ to be useful if

$$\frac{\bar{R}_{(n)}}{\bar{R}_{(n-1)}} \geq \tilde{\tau} \quad (56)$$

for a threshold $\tilde{\tau}$ above 1. We note that while any finite value of $\tilde{\tau}$ that is larger than and bounded away from 1 works in the asymptotic regime where $n \rightarrow \infty$, in the small column size regime, for each value of n , $\tilde{\tau}$ is a heuristically chosen threshold slightly larger than 1, *e.g.* $\tilde{\tau} = 1.5$ for $n = 100$.

After deciding that Φ is useful, the j^{th} retained column is inferred as:

$$\hat{I}_R(j) = \operatorname{argmax}_{i \in [n]} M_{i,j}(\Phi) \quad (57)$$

The resulting modified deletion detection algorithm is given in Algorithm 5. From Figure 5, it can be seen that Algorithm 5 performs well for a small column size.

Comparing the different curves in Figure 5, we can conclude that as the crossover probability ϵ increases, the decay of the error probability becomes slower, indicating a worse detection performance. In turn, to achieve the same deletion detection error rate, we need a higher number of seeds in scenarios with a higher level of obfuscation.

Algorithm 5 Modified Distribution-Agnostic Seeded Deletion Detection Algorithm

Input : $(\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \tilde{\tau})$
Output: \hat{I}_R ; /* Retention set */
 $\mathcal{S}(\mathcal{X}) \leftarrow \text{SymmetryGroup}(\mathcal{X})$;
/* Sweep over Φ until a useful remapping is found. */
for $s \leftarrow 1$ **to** $|\mathcal{X}|$! **do**
 $\hat{I}_R \leftarrow \emptyset$;
 $\Phi \leftarrow \mathcal{S}(\mathcal{X})[s]$; /* Pick a remapping. */
 $\mathbf{L}(\Phi) \leftarrow \text{HammDist}(\mathbf{G}^{(1)}, \mathbf{G}^{(2)}(\Phi))$; /* Eq. (18) */
 $\mu(\Phi) \leftarrow \text{SampleMean}(\mathbf{L}(\Phi))$; /* Eq. (28) */
 $\mathbf{M}(\Phi) \leftarrow |\mathbf{L}(\Phi) - \mu(\Phi)|$; /* Eq. (29) */
 $\mathbf{R} \leftarrow \text{OrderStatisticRatios}(\mathbf{M}(\Phi))$; /* Eq. (53) */
 if $\text{mean}(\mathbf{R}[n][:]) < \tilde{\tau} \text{ mean}(\mathbf{R}[n-1][:])$ **then**
 /* Φ is a useless remapping. */
 if $s \neq |\mathcal{X}|$! **then**
 Skip to next Φ ;
 else
 /* No useful remapping is found. */
 return ERROR ;
 end
 else
 /* Φ is a useful remapping. */
 /* Sweep over all $G_j^{(2)}(\Phi)$. */
 for $j \leftarrow 1$ **to** \tilde{K}_n **do**
 /* Find $G_i^{(1)}$ correlated with $G_j^{(2)}(\Phi)$. */
 /* Add such i to the retention set. */
 $\hat{I}_R \leftarrow \hat{I}_R \cup \{\underset{i \in [n]}{\text{argmax}} \mathbf{M}(\Phi)[i][j]\}$;
 end
 return \hat{I}_R ;
 end
end

C. De-Anonymization Scheme

In this subsection, we evaluate the performance of the de-anonymization scheme (Algorithm 3) proposed in Section III-C. While doing so, we decrease the column size from $n = 100$ to $n = 25$ due to the following considerations:

- (i) As discussed in Section III-C, Algorithm 3 has a runtime of $\mathcal{O}(m_n^2 n)$.
- (ii) As in the proof of Theorem 1 (Eq. (40)), the matching error probability decreases exponentially in the column size n . Hence to observe a non-trivial (non-zero) matching error, we must consider a row size m_n exponential in n (Definition 7). This suggests, simulating Algorithm 3 in the non-trivial regime becomes computationally prohibitive, even for $n = 100$.

In addition to the modified deletion detection algorithm (Algorithm 5), we make a slight modification to Algorithm 3, described below to accommodate the small column size. After performing replica and deletion detection, and estimating $\hat{p}_{X,Y^S|S}$, instead of fixing an $\epsilon > 0$ (that depends on n) and checking the ϵ -joint-typicality of the row pairs from \mathbf{X} and $\tilde{\mathbf{Y}}$, in Algorithm 6, we do the following:

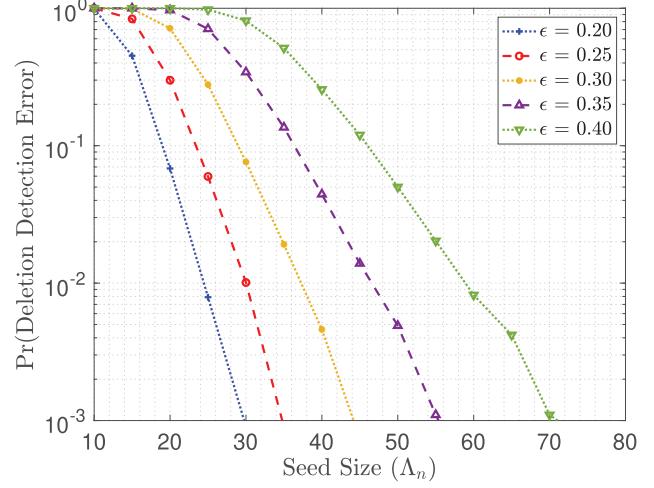


Fig. 5. Probability of error of the modified seeded deletion detection algorithm (Algorithm 5) vs. the seed size Λ_n with $\tilde{\tau} = 1.5$ and 10^4 trials. The y-axis is given in the logarithmic domain to validate the exponential relation between the deletion detection error probability and Λ_n given by Lemma 2. Different curves correspond to different crossover probabilities ϵ .

- (i) For each $(i, j) \in [m_n]^2$, compute

$$\Delta_{i,j} \triangleq |\hat{H}(X, Y^S|S) - \hat{H}_{i,j}(X, Y^S|S)| \quad (58)$$

where

$$\hat{H}_{i,j}(X, Y^S|S) \triangleq -\frac{1}{n} \log \hat{p}_{X^n, Y^{K_n}|S^n}(X_i^n, Y_j^{K_n}|\hat{S}^n) \quad (59)$$

- (ii) For each $j \in [m_n]$, match the j^{th} row $Y_j^{K_n}$ of $\tilde{\mathbf{Y}}$ with the l^{th} row X_l^n of \mathbf{X} , if

$$l = \arg \min_{i \in [m_n]} \Delta_{i,j} \quad (60)$$

assigning $\hat{\sigma}_n(l) = j$.

- (iii) If there exist distinct j_1, j_2 such that l satisfies (60), declare error and assign $\hat{\sigma}_n(l) = 0$.

We note that testing $\Delta_{i,j}$ against a threshold ϵ would correspond to the joint-typicality check as done in Algorithm 3.

The experimental evaluation of the performance of Algorithm 6 is given in Figure 6. As expected, we observe that a higher row size m_n leads to an increased matching error.

Similar to Figures 4 and 5, Figure 6 demonstrates the impact of the obfuscation on the de-anonymization performance. As the databases become more obfuscated, the de-anonymization performance degrades potentially an order of magnitude. Hence, we can conclude that the amount of obfuscation plays a crucial role in preserving privacy in databases.

D. Histogram-Based Repetition Detection

In this subsection, we evaluate the performance of the histogram-based repetition detection algorithm (Algorithm 4) of Section III-E in the no-obfuscation setting.

For uniform database distribution $X \sim \text{Unif}([\mathcal{X}])$, we can obtain the following closed-form asymptotic expression for the repetition detection error probability of Algorithm 4:

Algorithm 6 Modified Distribution-Agnostic De-Anonymization Scheme

Input : $(\mathbf{X}, \mathbf{Y}, \mathbf{G}^{(1)}, \mathbf{G}^{(2)})$

Output: $\hat{\sigma}_n$; */* Anonymization function */*

/ Noisy Replica Detection (Sec. II-A) */*

isReplica \leftarrow Alg.1(\mathbf{Y});

/ Modified Seeded Deletion Detection (Sec. IV-B) */*

$\mathbf{G}^{(2)} \leftarrow$ removeExtraReplicas($\mathbf{G}^{(2)}$, *isReplica*);

isDeleted \leftarrow Alg.5($\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \tilde{\tau}$);

$\hat{S}^n \leftarrow$ EstimateRepetitionPattern(*isReplica*, *isDeleted*);

/ Estimate the underlying distribution. */*

$\hat{p}_{X, Y^S|S} \leftarrow$ EstimateDist($\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \hat{S}^n$);

$\tilde{\mathbf{Y}} \leftarrow$ MarkerAddition(\mathbf{Y}, \hat{S}^n);

/ Computing $\Delta_{i,j}$ Eq. (58) */*

$\Delta \leftarrow$ computeDeltas($\mathbf{X}, \mathbf{Y}, \hat{S}^n, \hat{p}_{X, Y^S|S}$);

/ Sweep over all row pairs (X_i, Y_j) . */*

for $i = 1$ **to** *rowSize*(\mathbf{X}) **do**

count $\leftarrow 0$;

for $j = 1$ **to** *rowSize*($\tilde{\mathbf{Y}}$) **do**

if $i = \operatorname{argmin}_{l \in [m_n]} \Delta_{l,j}$ **then**

/ X_i and Y_j are matching rows. */*

$\hat{\sigma}_n[i] \leftarrow j$;

count \leftarrow *count* + 1;

end

end

if *count* $\neq 1$ **then**

/ X_i is matched with multiple Y_j . */*

$\hat{\sigma}_n[i] \leftarrow 0$; */* Matching error. */*

end

end

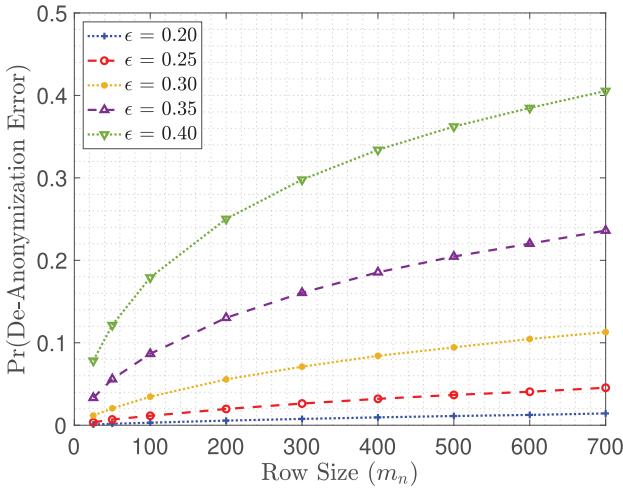


Fig. 6. Probability of de-anonymization error of the modified de-anonymization scheme (Algorithm 6) vs. the row size m_n with $n = 25$, $\Lambda_n = 25$ and 10^4 trials. Different curves correspond to different crossover probabilities ϵ .

Proposition 2 [15, Proposition 4]: Let ξ_n denote the probability of the column histograms of \mathbf{X} not being unique. If $\mathbf{X} \sim \text{Unif}(\mathcal{X})$, then

$$\xi_n = n^2 m_n^{\frac{1-|\mathcal{X}|}{2}} (4\pi)^{\frac{1-|\mathcal{X}|}{2}} |\mathcal{X}|^{\frac{|\mathcal{X}|}{2}} (1 + o_{m_n}(1))(1 - o_n(1)) \quad (61)$$

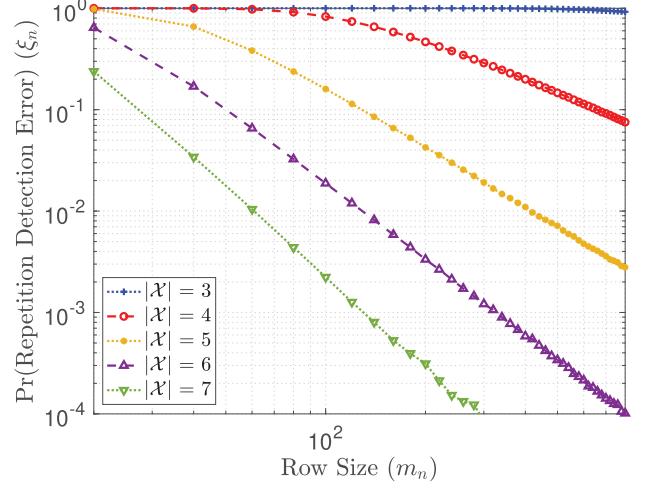


Fig. 7. Probability of error of the histogram-based repetition detection (Section III-E) vs. the row size m_n with 10^6 trials. Both axes are given in the logarithmic domain to validate the linear relation given in (62).

Proposition 2 states that for fixed n and $|\mathcal{X}|$, $\xi_n \propto m_n^{\frac{1-|\mathcal{X}|}{2}}$, i.e., $\log \xi_n$ is linear with $\log m_n$ for large n and m_n . More formally,

$$\log \xi_n = \frac{1-|\mathcal{X}|}{2} \log m_n + C_{|\mathcal{X}|} + 2 \log n + \zeta_{m_n, n} \quad (62)$$

where

$$C_{|\mathcal{X}|} \triangleq \frac{1-|\mathcal{X}|}{2} \log 4\pi + \frac{|\mathcal{X}|}{2} \log |\mathcal{X}| \quad (63)$$

and $\zeta_{m_n, n} \rightarrow 0$ as $m_n, n \rightarrow \infty$. We note that (62) implies that the order relation given in Lemma 3 is tight.

From Figure 7, one can see that after some m_n , $\log \xi_n$ decays linearly with $\log m_n$ (excluding $|\mathcal{X}| = 3$) with respective approximate slopes of $-1.40, -1.97, -2.51, -2.97$ demonstrating that (62), which predicts respective asymptotic slopes of $-1.50, -2, -2.50, -3$, holds even when $n = 100$. Note the value of m_n where ξ_n starts decaying becomes smaller with increasing alphabet size $|\mathcal{X}|$.

In Figure 7, as the alphabet \mathcal{X} gets larger, the decay of the error probability is steeper. This is inherent to the nature of the histograms as in Proposition 2 and not an artifact of the simulation configuration chosen.

V. CONCLUSION

In this paper, we have investigated the distribution-agnostic database de-anonymization problem under synchronization errors, in the form of column repetitions, and obfuscation, in the form of noise on the retained entries of the database. We showed that through either modifications of the existing algorithms for the distribution-agnostic setting or novel ones, we can correctly detect the repetition pattern and perform de-anonymization, in both obfuscated (noisy) and no-obfuscation (noiseless) settings. Interestingly, our results show that in terms of matching capacity, there is no penalty for not knowing the underlying distributions asymptotically. Our experimental results illustrate that our proposed algorithms or their slightly modified versions work in the non-asymptotic regime with small database sizes as well. Overall, our work provides insights into the practicality of distribution-agnostic

database de-anonymization algorithms with theoretical guarantees. Finally, an important direction of future work is the extension of our proposed algorithms to continuous database distributions, as well as adversarial noise and column repetition patterns.

APPENDIX

A. Proof of Lemma 1

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

$$F_k \triangleq \frac{1}{K_n - 1} \sum_{j=1}^{K_n-1} \prod_{i=0}^{k-1} \frac{W_j - i}{m_n - i}, \quad \forall k \in [m_n] \quad (64)$$

and let

$$U \triangleq \frac{F_3 - F_1 F_2}{F_2 - F_1^2} \quad (65)$$

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

$$\hat{p}_0 = \frac{U + \sqrt{U^2 - 4UF_1 + 4F_2}}{2} \quad (66)$$

$$\hat{p}_1 = \frac{U - \sqrt{U^2 - 4UF_1 + 4F_2}}{2} \quad (67)$$

From [21], we get $\hat{p}_i \xrightarrow{p} p_i$, $i = 0, 1$, and in turn $\tau \xrightarrow{p} \frac{U}{2}$. Thus for large n , τ is bounded away from p_0 and p_1 . We complete the proof following the same steps taken in the proof of [15, Lemma 1], which we provide below for the sake of completeness.

Let A_j denote the event that $C_j^{(2)}$ and $C_{j+1}^{(2)}$ are noisy replicas and B_j denote the event that the algorithm infers $C_j^{(2)}$ and $C_{j+1}^{(2)}$ as replicas. Via the union bound, we can upper bound the total probability of replica detection error $\kappa_n^{(1)}$ as

$$\kappa_n^{(1)} = \Pr\left(\bigcup_{j=1}^{K_n-1} E_j\right) \quad (68)$$

$$\leq \sum_{j=1}^{K_n-1} \Pr(A_j^c) \Pr(B_j|A_j^c) + \Pr(A_j) \Pr(B_j^c|A_j) \quad (69)$$

Observe that conditioned on A_j^c , $W_j \sim \text{Binom}(m_n, p_0)$ and conditioned on A_j , $W_j \sim \text{Binom}(m_n, p_1)$. Then, from the Chernoff bound [26, Lemma 4.7.2], we get

$$\Pr(B_j|A_j^c) \leq 2^{-m_n D(\tau||p_0)} \quad (70)$$

$$\Pr(B_j^c|A_j) \leq 2^{-m_n D(1-\tau||1-p_1)} \quad (71)$$

Thus, through the union bound, we obtain

$$\kappa_n^{(1)} \leq (K_n - 1) \left[2^{-m_n D(\tau||p_0)} + 2^{-m_n D(1-\tau||1-p_1)} \right] \quad (72)$$

Since the RHS of (72) has $2K_n - 2 = \mathcal{O}(n)$ terms decaying exponentially in m_n , for any $m_n = \omega(\log n)$ we have

$$\kappa_n^{(1)} \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (73)$$

Observing that $n \sim \log m_n$ concludes the proof. \square

B. Proof of Lemma 2

For now, suppose that Φ is a useful remapping. Using Chebyshev's inequality [27, Theorem 4.2] it is straightforward to prove that for any $\epsilon_n > 0$

$$\gamma \triangleq \Pr(|\mu(\Phi) - \Lambda_n q_0(\Phi)| > \Lambda_n \epsilon_n) \quad (74)$$

$$= \mathcal{O}\left(\frac{1}{\tilde{K}_n n \Lambda_n \epsilon_n}\right) \quad (75)$$

First, let

$$I_R = \{r_1, \dots, r_{\tilde{K}_n}\} \quad (76)$$

and note $L_{r_j, j}(\Phi) \sim \text{Binom}(\Lambda_n, q_1(\Phi))$. Thus, from the Chernoff bound [26, Lemma 4.7.2] we get

$$\beta_{r_j, j} \triangleq \Pr(|L_{r_j, j}(\Phi) - \Lambda_n q_1(\Phi)| \geq \epsilon_n \Lambda_n) \quad (77)$$

$$\leq 2^{-\Lambda_n D(q_1(\Phi) - \epsilon_n || q_1(\Phi))} + 2^{-\Lambda_n D(1 - q_1(\Phi) - \epsilon_n || 1 - q_1(\Phi))}. \quad (78)$$

Now, for notational brevity, let

$$f(\epsilon) \triangleq D(q - \epsilon || q) \quad (79)$$

Then, one can simply verify the following

$$f'(\epsilon) = \log \frac{q}{1-q} - \log \frac{q-\epsilon}{1-q-\epsilon} \quad (80)$$

$$f''(\epsilon) = \frac{1}{\log e} \left[\frac{1}{q-\epsilon} + \frac{1}{1-q+\epsilon} \right] \quad (81)$$

Similarly, letting

$$g(\epsilon) \triangleq D(1 - q - \epsilon || 1 - q) \quad (82)$$

we get

$$g'(\epsilon) = \log \frac{1-q}{q} - \log \frac{1-q-\epsilon}{q+\epsilon} \quad (83)$$

$$g''(\epsilon) = \frac{1}{\log e} \left[\frac{1}{1-q-\epsilon} + \frac{1}{q+\epsilon} \right] \quad (84)$$

Observing that

$$f(0) = f'(0) = 0 \quad (85)$$

$$g(0) = g'(0) = 0 \quad (86)$$

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

$$f(\epsilon) = c(q)\epsilon^2 + \mathcal{O}(\epsilon^3) \quad (87)$$

$$g(\epsilon) = c(q)\epsilon^2 + \mathcal{O}(\epsilon^3) \quad (88)$$

where

$$c(q) \triangleq \frac{1}{\log e} \left[\frac{1}{q} + \frac{1}{1-q} \right] \quad (89)$$

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

$$\hat{\tau}_n = 2\Lambda_n \epsilon_n = o_n(\Lambda_n) \quad (90)$$

$$\Lambda_n \epsilon_n^2 = \Gamma_n^{1/3} \log n = \omega_n(\log n) \quad (91)$$

Then, we have

$$\beta_{r_j, j} \leq 2^{1 - \Lambda_n (c(q_1(\Phi)) \epsilon_n^2 + \mathcal{O}(\epsilon_n^3))} \quad (92)$$

$$= 2^{1-c(q_1(\Phi))\Gamma_n^{1/3} \log n + \mathcal{O}(\epsilon_n^3)} \quad (93)$$

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

$$|\mu(\Phi) - \Lambda_n q_0(\Phi)| \leq \Lambda_n \epsilon_n \quad (94)$$

$$|L_{r_j,j}(\Phi) - \Lambda_n q_1(\Phi)| \geq \Lambda_n \epsilon_n \quad (95)$$

From the triangle inequality, we have

$$M_{r_j,j}(\Phi) = |L_{r_j,j}(\Phi) - \mu(\Phi)| \quad (96)$$

$$\geq \Lambda_n (|q_1(\Phi) - q_0(\Phi)| - 2\epsilon_n) \quad (97)$$

$$\geq \hat{\tau}_n \quad (98)$$

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

$$\begin{aligned} \Pr(\exists j \in [\tilde{K}_n] : M_{r_j,j}(\Phi) \leq \hat{\tau}_n) \\ \leq \gamma + \sum_{j=1}^{\tilde{K}_n} \beta_{r_j,j} \\ = \gamma + 2^{\log \tilde{K}_n - 1 - c(q_1(\Phi))\Gamma_n^{1/3} \log n + \mathcal{O}(\epsilon_n^3)} \end{aligned} \quad (99) \quad (100) \quad (101)$$

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

$$\lim_{n \rightarrow \infty} \log \tilde{K}_n - c(q_1(\Phi))\Gamma_n^{1/3} \log n = -\infty \quad (102)$$

Thus we have

$$\lim_{n \rightarrow \infty} \Pr(\exists j \in [\tilde{K}_n] : M_{r_j,j} \leq \hat{\tau}_n) = 0 \quad (103)$$

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

$$\beta_{i,j} \triangleq \Pr(|L_{i,j}(\Phi) - \Lambda_n q_0(\Phi)| \geq \epsilon_n \Lambda_n) \quad (104)$$

$$\leq 2^{-\Lambda_n D(q_0(\Phi) - \epsilon_n \| q_0(\Phi))} \quad (105)$$

$$+ 2^{-\Lambda_n D(1 - q_0(\Phi) - \epsilon_n \| 1 - q_0(\Phi))} \quad (106)$$

$$= 2^{1-c(q_0(\Phi))\Gamma_n^{1/3} \log n + \mathcal{O}(\epsilon_n^3)} \quad (107)$$

Again, from the triangle inequality, we get

$$M_{i,j}(\Phi) = |L_{i,j}(\Phi) - \mu(\Phi)| \quad (108)$$

$$\leq 2\epsilon_n \quad (109)$$

$$= \hat{\tau}_n \quad (109)$$

From the union bound, we obtain

$$\Pr(\exists j \in [\tilde{K}_n] \exists i \in [n] \setminus \{r_j\} : M_{i,j}(\Phi) \geq \hat{\tau}_n) \quad (110)$$

$$\leq \gamma + \sum_{j=1}^{\tilde{K}_n} \sum_{i \neq r_j} \beta_{i,j} \quad (111)$$

$$\leq \gamma + n^2 2^{1-c(q_0(\Phi))\Gamma_n^{1/3} \log n + \mathcal{O}(\epsilon_n^3)} \quad (112)$$

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

$$\Pr(\exists j \in [\tilde{K}_n] \exists i \in [n] \setminus \{r_j\} : M_{i,j}(\Phi) \geq \hat{\tau}_n) \rightarrow 0. \quad (113)$$

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

For any remapping Φ that is not useful, following the same steps, one can prove that

$\Pr(\text{Remapping } \Phi \text{ is declared useful, even though it is not.})$

$$\leq \gamma + \sum_{i=1}^n \sum_{j=1}^{\tilde{K}_n} \Pr(M_{i,j} \geq \epsilon_n \Lambda_n) \quad (114)$$

$$\leq \gamma + n^2 2^{1-c(q_0(\Phi))\Gamma_n^{1/3} \log n + \mathcal{O}(\epsilon_n^3)} \quad (115)$$

$$= o_n(1) \quad (116)$$

Since $|\mathcal{S}(\mathcal{X})| = |\mathcal{X}|! = \mathcal{O}_n(1)$, we have

$$\sum_{\Phi: \text{not useful}} \Pr(\Phi \text{ is declared useful.}) = o_n(1) \quad (117)$$

concluding the proof. \square

C. Proof of Theorem 2

Let $\epsilon > 0$ and denote by ξ_n the error probability of the histogram-based repetition algorithm. Denote the ϵ -typical set of sequences (with respect to p_X) of length $k = (1 - \delta - \epsilon)n$ by $A_\epsilon^{(k)}(X)$ and the pairwise collision probability between X_1^n and X_i^n , given $\hat{K}_n = k$ and $X_1^n \in A_\epsilon^{(k)}(X)$, by $P_{col,i}(k)$. Since additional columns in $\bar{\mathbf{Y}}$ would decrease the pairwise collision probability between independent rows, we have

$$P_{col,i}(\hat{K}_n) \leq P_{col,i}(k), \quad \forall \hat{K}_n \geq k \quad (118)$$

Since we perform exact row matching and the rows of $\bar{\mathbf{X}}$ are independent, we have

$$P_{col,i}(k) = \Pr(\bar{X}_i^k = \bar{X}_1^k | \bar{X}_1^k \in A_\epsilon^{(k)}(X)) \quad (119)$$

$$\leq 2^{-k(H(X) - \epsilon)}. \quad (120)$$

Thus, we can bound the probability of error P_e as

$$P_e \leq \sum_{i=2}^{m_n} P_{col,i}(k) + \epsilon + \kappa_n + \xi_n \quad (121)$$

$$\leq 2^{nR} 2^{-k(H(X) - \epsilon)} + \epsilon + \kappa_n + \xi_n \quad (122)$$

where $\kappa_n = \Pr(\hat{K}_n < k)$. Since m_n is exponential in n , by Lemma 3, $\xi_n \rightarrow 0$ as $n \rightarrow \infty$. Furthermore, \hat{K}_n is a $\text{Binom}(n, 1 - \delta)$ random variable and from law of large numbers $\kappa_n \rightarrow 0$ as $n \rightarrow \infty$. Thus $P_e \leq \epsilon$ as $n \rightarrow \infty$ if

$$R < (1 - \delta - \epsilon)H(X). \quad (123)$$

Thus, we can argue that any database growth rate R satisfying

$$R < (1 - \delta)H(X) \quad (124)$$

is achievable, by taking ϵ small enough. \square

REFERENCES

- [1] P. Ohm, "Broken promises of privacy: Responding to the surprising failure of anonymization," *UCLA Law Rev.*, vol. 57, p. 1701, Aug. 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, Feb. 2016.
- [3] A. Datta, D. Sharma, and A. Sinha, "Provable de-anonymization of large datasets with sparse dimensions," in *Proc. Int. Conf. Princ. Secur. Trust.* Cham, Switzerland: Springer, Jan. 2012, pp. 229–248.
- [4] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Security Privacy*, May 2008, pp. 111–125.
- [5] L. Sweeney, "Weaving technology and policy together to maintain confidentiality," *J. Law, Med. Ethics*, vol. 25, nos. 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 Trans. Inf. Theory*, vol. 65, no. 2, pp. 724–741, Feb. 2019.
- [7] D. Cullina, P. Mittal, and N. Kiyavash, “Fundamental limits of database alignment,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 651–655.
- [8] F. Shirani, S. Garg, and E. Erkip, “A concentration of measure approach to database de-anonymization,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 2748–2752.
- [9] O. E. Dai, D. Cullina, and N. Kiyavash, “Database alignment with Gaussian features,” in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, Apr. 2019, pp. 3225–3233.
- [10] D. Kunisky and J. Niles-Weed, “Strong recovery of geometric planted matchings,” in *Proc. Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*, Jan. 2022, pp. 834–876.
- [11] R. Tamir, “On correlation detection of Gaussian databases via local decision making,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2023, pp. 1231–1236.
- [12] S. Bakirtas and E. Erkip, “Database matching under column deletions.” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 2720–2725.
- [13] S. Bakirtas and E. Erkip, “Matching of Markov databases under random column repetitions,” in *Proc. 56th Asilomar Conf. Signals, Syst., Comput.*, Oct. 2022, pp. 1248–1254.
- [14] S. Bakirtas and E. Erkip, “Seeded database matching under noisy column repetitions,” in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2022, pp. 386–391.
- [15] S. Bakirtas and E. Erkip, “Database matching under noisy synchronization errors,” *IEEE Trans. Inf. Theory*, vol. 70, no. 6, pp. 4335–4367, Jun. 2024.
- [16] S. Bakirtas and E. Erkip, “Database matching under adversarial column deletions,” in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr. 2023, pp. 181–185.
- [17] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. Cambridge, MA, USA: MIT Press, 2022.
- [18] T. M. Cover, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2006.
- [19] F. Shirani, S. Garg, and E. Erkip, “A concentration of measure approach to correlated graph matching,” *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, pp. 338–351, Mar. 2021.
- [20] G. Morvai and B. Weiss, “Order estimation of Markov chains,” *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1496–1497, Apr. 2005.
- [21] W. R. Blischke, “Moment estimators for the parameters of a mixture of two binomial distributions,” *Ann. Math. Statist.*, vol. 33, no. 2, pp. 444–454, Jun. 1962.
- [22] D. S. Moore and S. Kirkland, *The Basic Practice of Statistics*, vol. 2. New York, NY, USA: WH Freeman, 2007.
- [23] T. W. Anderson and L. A. Goodman, “Statistical inference about Markov chains,” *Ann. Math. Statist.*, vol. 28, no. 1, pp. 89–110, Mar. 1957.
- [24] A. W. Van der Vaart, *Asymptotic Statistics*, vol. 3. Cambridge, U.K.: Cambridge Univ. Press, 2000.
- [25] N. Balakrishnan and A. Stepanov, “Asymptotic properties of the ratio of order statistics,” *Statist. Probab. Lett.*, vol. 78, no. 3, pp. 301–310, Feb. 2008.
- [26] R. B. Ash, *Information Theory*. North Chelmsford, MA, USA: Courier Corporation, 2012.
- [27] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, vol. 26. Cham, Switzerland: Springer, 2004.