Database Matching Under Column Repetitions

Serhat Bakirtas, Elza Erkip

NYU Tandon School of Engineering

Emails: {serhat.bakirtas, elza}@nyu.edu

Abstract—Motivated by synchronization errors in the sampling of time-indexed databases, matching of random databases under random column repetitions (including deletions) is investigated. Column histograms are used as a permutation-invariant feature to detect the repetition pattern, whose asymptotic-uniqueness is proved using information-theoretic tools. Repetition detection is followed by a row matching scheme. Considering this overall scheme, sufficient conditions for successful database matching in terms of the database growth rate are derived. A modified version of Fano's inequality leads to a tight necessary condition for successful matching, establishing the matching capacity under column repetitions. This capacity is equal to the erasure bound, which assumes the repetition locations are known a-priori.

I. Introduction

In the past decades, collection of personal data has accelerated, mostly due to the proliferation of smart devices and the emergence of big data applications. Collected data is sold or published by companies and public institutions for commercial or academic purposes. Although data is published typically after *anonymization*, which refers to removing the explicit identifiers such as names and social security numbers, there has been a growing concern over potential privacy leakage, from legal [1] and corporate [2] points of view. In [3]–[7], using real data, it was shown that anonymization by itself cannot prevent the privacy leakage. However, these works lack a systematic and comprehensive understanding of the conditions under which the databases are prone to privacy attacks.

Recently the de-anonymization of correlated databases have been investigated from an information-theoretic perspective [8]–[10]. In [8], Shirani *et al.* considered a pair of databases of the same size and drew an analogy between channel decoding and database matching to derive necessary and sufficient conditions on the database growth rate for reliable database matching. In [9], Cullina *et al.* introduced *cycle mutual information* as a correlation metric and allowing arbitrary column sizes, derived sufficient conditions for successful matching and a converse result, using an error criterion different than [8].

In [10], motivated by synchronization errors in sampling of time-series datasets, we considered database matching under *random column deletions*. We assumed two databases with the same number of users (rows), generated according to a bivariate stochastic process, similar to [8], [9]. However, unlike [8], [9], in our model, the two databases have different number of attributes (columns), caused by unknown random
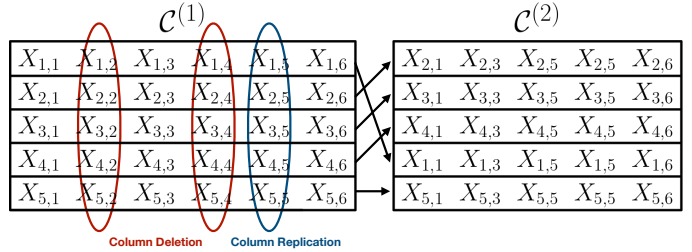
Fig. 1. An illustrative example of database matching under column repetitions. The columns circled in red are deleted whereas the column circled in blue is repeated twice, *i.e.*, replicated. Our goal is to estimate the row permutation $\Theta = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 1 & 2 & 3 & 5 \end{pmatrix}$, by matching the rows of $\mathscr{C}^{(1)}$ and $\mathscr{C}^{(2)}$. Here the ith row of $\mathscr{C}^{(1)}$ corresponds to the $\Theta(i)^{\text{th}}$ row of $\mathscr{C}^{(2)}$.

column deletions. Assuming the same error criterion as [8] and utilizing tools from deletion channel literature [11] we derived an achievable database growth rate assuming a probabilistic side information on the deletion locations. We also proposed an algorithm to extract the side information on the deletion locations from a batch of already-matched rows, called *seeds*.

In this paper, we generalize [10], by assuming a more general type of synchronization errors, namely *column repetitions*. In our model, in addition to some columns being deleted, some columns may be sampled several times consecutively, *i.e.*, *replicated*, as illustrated in Figure 1. Under this generalized model, we derive an improved achievable database growth rate. Our matching scheme uses a *histogram-based repetition detection* algorithm, which exploits the asymptotically large size of the database and the identical nature of the repetition across the rows of the database to infer the repetition pattern with high probability, rendering the seeds unnecessary. Furthermore, we derive the necessary conditions for successful matching. We show that the necessary and sufficient conditions are tight up to equality, and equal to the erasure bound, which is obtained when there are no replications and the deletion locations are perfectly known. Thus, we characterize the *capacity* of the matching of column repeated databases.

In a companion paper [12], we investigate database matching under column repetitions as well as noise. As shown in [12], noise renders histogram-based repetition algorithm insufficient, and necessitates seeds for successful matching. In addition, although the deletion and replica detection algorithms of [12] work in the noiseless setting, unlike the algorithms presented here, they still require seeds.

The organization of this paper is as follows: Section II

contains the problem formulation. In Section III, our main result on matching capacity and the proof of the achievability are presented. In Section IV, the converse is proved. Finally, in Section V the results and ongoing work are discussed.

Notation: We denote the set of integers $\{1, 2, ..., n\}$ as $[n]$, databases with calligraphic letters, random vectors with bold uppercase letters. $\mathbb{1}_\varepsilon$ denotes the indicator of event ε. The logarithms, unless stated explicitly, are in base 2.

II. PROBLEM FORMULATION

We use the following definitions, some of which are similar to [8], [10] to formalize our problem.

Definition 1. (Unlabeled Database) An (m, n, p_X) *unlabeled database* is a randomly generated $m \times n$ matrix $\mathscr{C} = \{X_{i,j} \in \mathfrak{X}^{m \times n}\}$ with *i.i.d.* entries drawn according to the distribution p_X with a finite support $\mathfrak{X} = \{1, \dots, |\mathfrak{X}|\}$.

For fixed column size n, as the row size m increases, the probability of matching error increases since there are more candidates. In the following definition, we define *database growth rate* characterizing the relationship between m and n for successful matching.

Definition 2. (Database Growth Rate) The *database growth rate* R of an (m, n, p_X) unlabeled database is defined as

$$R = \lim_{n \to \infty} \frac{1}{n} \log m \qquad (1)$$

Definition 3. (Column Repetition Pattern) The *column repetition pattern* $\mathbf{S}^n = \{S_1, S_2, ..., S_n\}$ is a random vector consisting of n *i.i.d.* entries drawn from a discrete probability distribution p_S with a finite discrete support $\{0, \dots, s_{\max}\}$. the parameter $\delta \triangleq p_S(0)$ is called the *deletion probability*.

Definition 4. (Labeled Repeated Database) Let $\mathscr{C}^{(1)}$ be an (m, n, p_X) unlabeled database, $\mathbf{S}^n$ be the repetition pattern, and Θ be a uniform permutation of $[m]$, independent of $\mathscr{C}^{(1)}$. Given $\mathscr{C}^{(1)}$ and $\mathbf{S}^n$, the pair $(\mathscr{C}^{(2)}, \Theta)$ is called the *labeled repeated database* if the ith column $\mathbf{R}_i^{(1)}$ of $\mathscr{C}^{(1)}$ and its counterpart $\mathbf{R}_i^{(2)}$ in $\mathscr{C}^{(2)}$ have the following relation:

$$\mathbf{R}_i^{(2)} = \begin{cases} E, & \text{if } S_i = 0 \\ (\Theta \circ \mathbf{R}_i^{(1)}) \otimes \mathbf{1}_{1 \times S_i} & \text{if } S_i \geq 1 \end{cases} \qquad (2)$$

where $\mathbf{1}_{1 \times S_i}$ and $\otimes$ denote the all-ones row vector of length S_i and the Kronecker product, respectively. Furthermore $\mathbf{R}_i^{(2)} = E$ corresponds to the empty string and $\mathbf{R}_i^{(2)} = (\Theta \circ \mathbf{R}_i^{(1)}) \otimes \mathbf{1}_{1 \times S_i}$ corresponds to $\mathbf{R}_i^{(2)}$ being an $m \times S_i$ matrix consisting of S_i copies of $\Theta \circ \mathbf{R}_i^{(1)}$, concatenated together.

For the labeled and unlabeled databases in Definition 4, the ith row $\mathbf{Y}_i$ of $\mathscr{C}^{(2)}$ is said to correspond to the user $\Theta^{-1}(i)$. The rows $\mathbf{X}_{i_1}$ and $\mathbf{Y}_{i_2}$ are said to be *matching rows*, if $\Theta(i_1) = i_2$, where Θ is called the *labeling function*.

When $S_i = 0$, the ith column of $\mathscr{C}^{(1)}$ is said to be *deleted* and when $S_i > 1$, i^{th} column of $\mathscr{C}^{(1)}$ is said to be *replicated*.

In this paper we assume that $\mathbf{S}^n$ and $\mathscr{C}^{(1)}$ are independent. In our model, $\mathscr{C}^{(2)}$ is obtained by shuffling $\mathscr{C}^{(1)}$ with Θ

followed by column repetition, and there is no noise on the retained entries, similar to the repeat channel model [13]. In our companion paper [12], we investigate the repetition setup with noise, where entrywise noise follows the column repetition.

Definition 5. (Successful Matching Scheme) A *matching scheme* is a sequence of mappings $\phi_n : (\mathscr{C}^{(1)}, \mathscr{C}^{(2)}) \to \hat{\Theta}_n$ where the permutation $\hat{\Theta}_n$ of $[m]$, is the estimate of the correct labeling function Θ_n. The scheme ϕ_n is *successful* if

$$\Pr\left(\Theta_n(I) = \hat{\Theta}_n(I)\right) \to 1 \text{ as } n \to \infty \qquad (3)$$

where the index I is drawn uniformly from $[m]$.

Definition 6. (Achievable Database Growth Rate) Given a database probability distribution p_X and a repetition probability distribution p_S, a database growth rate R is said to be *achievable* if for any pair of databases $(\mathscr{C}^{(1)}, \mathscr{C}^{(2)})$ with these parameters, there exists a successful matching scheme.

Definition 7. (Matching Capacity) Given a database probability distribution p_X and a repetition probability distribution p_S, the *matching capacity C* is the supremum of the set of all achievable rates.

In this work, our goal is to characterize the matching capacity of the database matching problem under the random column repetition model presented above.

III. MAIN RESULT AND ACHIEVABILITY

Theorem 1 below presents our main result on the matching capacity. We prove the achievability part of Theorem 1 in this section and the converse in Section IV.

Theorem 1. *(Matching Capacity Under Column Repetitions) Consider an (m, n, p_X) unlabeled database and an independent column repetition pattern $\mathbf{S}^n$ with repetition distribution p_S. Then, the* matching capacity *is*

$$C = (1 - \delta)H(X) \qquad (4)$$

where $\delta \triangleq p_S(0)$ is the deletion probability and H is the entropy function.

Theorem 1 states that when there is no noise on the retained entries, we can achieve the erasure bound which assumes a-priori knowledge of the repetition pattern. Another observation is that the matching capacity only depends on the repetition distribution through the deletion probability $\delta = p_S(0)$. This renders the replicas irrelevant in the noiseless scenario, as we discuss in more detail in Section IV.

To prove the achievability, we consider the following two-phase matching scheme: Given $\mathscr{C}^{(1)}$ and $\mathscr{C}^{(2)}$, the labeled and the unlabeled databases, we first infer the underlying repetition pattern $\mathbf{S}^n$ using the *histogram-based* detection algorithm described below. Then, we use the rowwise matching scheme proposed in [10].

Let $\mathbf{H}_j^{(1)}$ denote the *histogram* of the jth column of $\mathscr{C}^{(1)}$. More formally,

$$\mathbf{H}_j^{(1)}(r) \triangleq \sum_{t=1}^{m} \mathbb{1}_{\left[\mathscr{C}_{t,j}^{(1)} = r\right]}, \forall r \in \mathfrak{X}, \forall j \in [n] \qquad (5)$$

Our histogram-based detection algorithm exploits two facts: First, the histogram and equivalently the type of each column of $\mathscr{C}^{(1)}$ and $\mathscr{C}^{(2)}$ are permutation-invariant. Second, as we prove in Lemma 1, the histogram of each column is asymptotically-unique due to the row size being exponential in the column size. Finally, since there is no noise on the retained entries of $\mathscr{C}^{(2)}$, we can match the histograms of the columns of $\mathscr{C}^{(1)}$ and $\mathscr{C}^{(2)}$, in an error-free fashion, provided that the histograms are unique.

The following lemma provides conditions for the asymptotic-uniqueness of the histograms.

Lemma 1. *(Asymptotic Uniqueness of the Histograms) Let* $\mathbf{H}_j^{(1)}$ *denote the histogram of the jth column of $\mathscr{C}^{(1)}$. Then,*

$$\Pr\left(\exists i, j \in [n], i \neq j, \mathbf{H}_i^{(1)} = \mathbf{H}_j^{(1)}\right) \to 0 \ as \ n \to \infty \qquad (6)$$

if $m = \omega(n^{\frac{4}{|\mathfrak{X}|-1}})$.

Proof. For brevity, let $\mu_n \triangleq \Pr(\exists i, j \in [n], i \neq j, \mathbf{H}_i^{(1)} = \mathbf{H}_j^{(1)})$. Notice that since the entries of $\mathscr{C}^{(1)}$ are *i.i.d.*, $\mathbf{H}_i^{(1)}$ are *i.i.d.* Multinomial(m, p_X) random variables. Then,

$$\mu_n \leq n^2 \Pr(\mathbf{H}_1^{(1)} = \mathbf{H}_2^{(1)}) \qquad (7)$$

$$= n^2 \sum_{\mathbf{h}} \Pr(\mathbf{H}_1^{(1)} = \mathbf{h})^2 \qquad (8)$$

where the sum is over all vectors of length $|\mathfrak{X}|$, summing up to m. Let $m_i \triangleq \mathbf{h}(i) \ i = 1, \ldots, |\mathfrak{X}|$. Then,

$$\Pr(\mathbf{H}_1^{(1)} = \mathbf{h}) = \binom{m}{m_1, m_2, \ldots, m_{|\mathfrak{X}|}} \prod_{i=1}^{|\mathfrak{X}|} p_X(i)^{m_i} \qquad (9)$$

Hence, we have

$$\mu_n \leq n^2 \sum_{m_1 + \cdots + m_{|\mathfrak{X}|} = m} \binom{m}{m_1, m_2, \ldots, m_{|\mathfrak{X}|}}^2 \prod_{i=1}^{|\mathfrak{X}|} p_X(i)^{2m_i} \qquad (10)$$

where $\binom{m}{m_1, m_2, \ldots, m_{|\mathfrak{X}|}}$ is the multinomial coefficient corresponding to the $|\mathfrak{X}|$-tuple $(m_1, \ldots, m_{|\mathfrak{X}|})$ and the summation is over all possible non-negative indices $m_1, \ldots, m_{|\mathfrak{X}|}$ whose sum is m.

From [14, Theorem 11.1.2], we have

$$\prod_{i=1}^{|\mathfrak{X}|} p_X(i)^{2m_i} = 2^{-2m(H(\tilde{p}) + D(\tilde{p}\|p_X))} \qquad (11)$$

where $\tilde{p}$ is the type corresponding to $|\mathfrak{X}|$-tuple $(m_1, \ldots, m_{|\mathfrak{X}|})$:

$$\tilde{p} = \left(\frac{m_1}{m}, \ldots, \frac{m_{|\mathfrak{X}|}}{m}\right) \qquad (12)$$

From Stirling's approximation, we get

$$\binom{m}{m_1, m_2, \ldots, m_{|\mathfrak{X}|}}^2 \leq e^2 (2\pi)^{-|\mathfrak{X}|} m^{1-|\mathfrak{X}|} \Pi_{\tilde{p}}^{-1} 2^{2mH(\tilde{p})} \qquad (13)$$

where $\Pi_{\tilde{p}} = \prod_{i=1}^{|\mathfrak{X}|} \tilde{p}(i)$.

Combining (10)-(13), we get

$$\mu_n \leq e^2 (2\pi)^{-|\mathfrak{X}|} n^2 m^{1-|\mathfrak{X}|} \sum_{\tilde{p}} \Pi_{\tilde{p}}^{-1} 2^{-2mD(\tilde{p}\|p_X)} \qquad (14)$$

Let

$$T = \sum_{\tilde{p}} \Pi_{\tilde{p}}^{-1} 2^{-2mD(\tilde{p}\|p_X)} = T_1 + T_2 \qquad (15)$$

where

$$T_1 = \sum_{\tilde{p}: D(\tilde{p}\|p_X) > \frac{1}{2\log_e 2}\varepsilon^2} \Pi_{\tilde{p}}^{-1} 2^{-2mD(\tilde{p}\|p_X)} \qquad (16)$$

$$T_2 = \sum_{\tilde{p}: D(\tilde{P}\|p_X) \leq \frac{1}{2\log_e 2}\varepsilon^2} \Pi_{\tilde{p}}^{-1} 2^{-2mD(\tilde{p}\|p_X)}, \qquad (17)$$

$D(\tilde{p}\|p_X)$ denotes the Kullback-Leibler divergence between $\tilde{p}$ and p_X, and ε, which is described below in more detail, is a small positive number decaying with n.

First, we look at T_2. From Pinsker's inequality [14, Lemma 11.6.1], we have

$$D(\tilde{p}\|p_X) \leq \frac{1}{2\log_e 2}\varepsilon^2 \implies \mathrm{TV}(\tilde{p}, p_X) \leq \varepsilon \qquad (18)$$

where TV denotes the total variation distance. Therefore

$$\left|\{\tilde{p}: D(\tilde{p}\|p_X) \leq \frac{1}{2\log_e 2}\varepsilon^2\}\right| \leq |\{\tilde{p}: \mathrm{TV}(\tilde{p}, p_X) \leq \varepsilon\}|$$
$$= O(m^{|\mathfrak{X}|-1}\varepsilon^{|\mathfrak{X}|-1}) \qquad (19)$$

where the last equality follows from the fact in a type we have $|\mathfrak{X}| - 1$ degrees of freedom, since the sum of the $|\mathfrak{X}|$-tuple $(m_1, \ldots, m_{|\mathfrak{X}|})$ is fixed. Furthermore, when $\mathrm{TV}(\tilde{p}, p_X) \leq \varepsilon$, we have

$$\Pi_{\tilde{p}} \geq \prod_{i=1}^{|\mathfrak{X}|} (p_X(i) - \varepsilon) \geq \Pi_{p_X} - \varepsilon \sum_{i=1}^{|\mathfrak{X}|} \prod_{j \neq i} p_X(j) \qquad (20)$$

Hence

$$\Pi_{\tilde{p}}^{-1} \leq \frac{1}{\Pi_{p_X} - \varepsilon \sum_{i=1}^{|\mathfrak{X}|} \prod_{j \neq i} p_X(j)} \qquad (21)$$

and

$$T_2 \leq \frac{1}{\Pi_{p_X} - \varepsilon \sum_{i=1}^{|\mathfrak{X}|} \prod_{j \neq i} p_X(j)} O(m^{|\mathfrak{X}|-1}\varepsilon^{|\mathfrak{X}|-1}) \qquad (22)$$

$$= O(m^{|\mathfrak{X}|-1}\varepsilon^{|\mathfrak{X}|-1}) \qquad (23)$$

for small ε.

Now, we look at T_1. Note that since $m_i \in \mathbb{Z}_+$, we have $\Pi_{\tilde{p}} \leq m^{|\mathfrak{X}|}$, suggesting the multiplicative term in the summation in (16) is polynomial with m. If $m_i = 0$ we can simply discard it and return to Stirling's approximation with the reduced number of categories. Furthermore, from [14, Theorem 11.1.1], we have

$$\left|\{\tilde{p}: D(\tilde{p}\|p_X) > \frac{1}{2\log_e 2}\varepsilon^2\}\right| \leq |\{\tilde{p}\}| \leq (m+1)^{|\mathfrak{X}|} \qquad (24)$$

suggesting the number of terms which we take the summation over in (16) is polynomial with m as well. Therefore, as long as $m\varepsilon^2 \to \infty$, T_1 has a polynomial number of elements which decay exponentially with m. Thus

$$T_1 \to 0 \text{ as } n \to \infty \qquad (25)$$

Define

$$U_i = e^2(2\pi)^{-|\mathfrak{X}|}m^{1-|\mathfrak{X}|}T_i, \ i = 1,2 \qquad (26)$$

and choose $\varepsilon = m^{-\frac{1}{2}}V$ for some V satisfying $V = \omega(1)$ and $V = o(m^{1/2})$. Thus, U_1 vanishes exponentially fast since $m\varepsilon^2 = V^2 \to \infty$ and

$$U_2 = O(\varepsilon^{|\mathfrak{X}|-1}) = O(m^{(1-|\mathfrak{X}|)/2}V^{(|\mathfrak{X}|-1)}). \qquad (27)$$

Combining (25)-(27), we have

$$U = U_1 + U_2 = O(m^{(1-|\mathfrak{X}|)/2}V^{(|\mathfrak{X}|-1)}) \qquad (28)$$

and we get

$$\mu_n \le n^2 O(m^{(1-|\mathfrak{X}|)/2}V^{(|\mathfrak{X}|-1)}) \qquad (29)$$

By the assumption $m = \omega(n^{\frac{4}{|\mathfrak{X}|-1}})$, we have $m = n^{\frac{4}{|\mathfrak{X}|-1}}W$ for some W satisfying $\lim_{n\to\infty} W = \infty$. Now, taking $V = o(W^{1/2})$ (e.g. $V = W^{1/3}$), we get

$$\mu_n \le O(n^2 n^{-2}W^{(1-|\mathfrak{X}|)/2}V^{(|\mathfrak{X}|-1)}) = o(1) \qquad (30)$$

Thus $m = \omega(n^{\frac{4}{|\mathfrak{X}|-1}})$ is enough to have $\mu_n \to 0$ as $n \to \infty$. $\square$

By Law of Large Numbers (LLN) $\mathbf{H}_i^{(1)}$ are very close to one another with high probability. However, Lemma 1 shows that, a row size $m = \omega(n^{\frac{4}{|\mathfrak{X}|-1}})$ provides enough granularity to guarantee the asymptotic-uniqueness of $\mathbf{H}_i^{(1)}$. Therefore, since m is exponential in n, the histograms of the columns of $\mathscr{C}^{(1)}$ are asymptotically-unique, allowing us to do histogram-based detection of deletions and replicas with a vanishing error probability.

Next, we present the proof of the achievability part of Theorem 1.

Proof of Achievability of Theorem 1. We need to show that for a given pair of matching rows, WLOG $\mathbf{X}_1$ of $\mathscr{C}^{(1)}$ and $\mathbf{Y}_l$ of $\mathscr{C}^{(2)}$, the probability of error $P_e \triangleq \Pr(\hat{\Theta}(1) \ne l)$ of the following matching scheme can be made arbitrarily small asymptotically.

Let $\mathbf{S}^n$ be the underlying repetition pattern, $K = \sum_{i=1}^n S_i$ denote the number of columns in $\mathscr{C}^{(2)}$. Our matching scheme consists of the following steps:

1) Construct the histogram matrices $\mathscr{H}^{(1)}$ and $\mathscr{H}^{(2)}$ as

$$\mathscr{H}_{i,j}^{(r)} = \sum_{t=1}^m \mathbb{1}_{\left[\mathscr{C}_{t,j}^{(r)}=i\right]}, \begin{cases} \forall j \in [n], \forall i \in \mathfrak{X} & \text{if } r = 1 \\ \forall j \in [K], \forall i \in \mathfrak{X} & \text{if } r = 2 \end{cases} \qquad (31)$$

2) Check whether the columns $\mathbf{H}_j^{(1)}$ $j \in [n]$ of $\mathscr{H}^{(1)}$ are unique. If there is at least one pair of identical columns, we declare a *detection error* whose probability is denoted by μ_n. Otherwise, proceed with the next step.
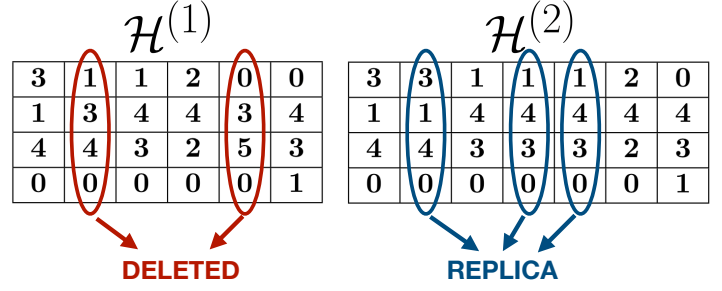

Fig. 2. An illustrative example of histogram-based detection with $|\mathfrak{X}| = 4$. The ith row denotes the occurrences of the ith symbol from $\mathfrak{X}$ in their respective columns, $\forall i \in \mathfrak{X}$. The columns circled in red are detected as deleted whereas the columns circled in blue are detected as extra replicas. Here $n = 6$, $K = 7$ and $\hat{K} = 4$ are the respective column sizes of $\mathscr{C}^{(1)}$, $\mathscr{C}^{(2)}$, and $\tilde{\mathscr{C}}^{(1)}$, as described in Steps 1-5.

3) If the jth column $\mathbf{H}_j^{(1)}$ of $\mathscr{H}^{(1)}$ is absent from $\mathscr{H}^{(2)}$, declare it deleted. Note that, provided the uniqueness of the histograms $\mathbf{H}_j^{(1)}$ $\forall j \in [n]$ (no error in Step 2), this step is error free.

4) If the jth column $\mathbf{H}_j^{(2)}$ of $\mathscr{H}^{(2)}$ is present multiple times consecutively, declare it replicated. Again, provided the uniqueness of the histograms $\mathbf{H}_j^{(1)}$ $\forall j \in [n]$ (no error in Step 2), this step is error free.

5) Discard the deleted columns from $\mathscr{C}^{(1)}$ and all-but-one of the replicas from $\mathscr{C}^{(2)}$, to obtain $\tilde{\mathscr{C}}^{(1)}$ and $\tilde{\mathscr{C}}^{(2)}$, respectively. Note that $\tilde{\mathscr{C}}^{(1)}$ and $\tilde{\mathscr{C}}^{(2)}$ have the same number of columns, $\hat{K} = n - \sum_{i=1}^n \mathbb{1}_{[S_i=0]}$.

6) Fix $\varepsilon > 0$. If $\hat{K} < k = n(1 - \delta - \varepsilon)$, declare error, whose probability is denoted by κ_n. Otherwise, proceed with the next step.

7) Match the lth row $\mathbf{Y}_l$ of $\tilde{\mathscr{C}}^{(2)}$ with the ith row $\mathbf{X}_1$ of $\tilde{\mathscr{C}}^{(1)}$, if $\mathbf{X}_1$ is the only ε-typical [14, Chapter 3] row of $\tilde{\mathscr{C}}^{(1)}$, equal to $\mathbf{Y}_l$, assigning $\hat{\Theta}(1) = l$. Otherwise, declare *collision error*. Note that there are $2^{nR} - 1$ potential collisions.

Denote the ε-typical set of sequences (with respect to p_X) of length k by $A_\varepsilon^{(k)}(X)$ and the pairwise collision probability between $\mathbf{X}_1$ and $\mathbf{X}_i$, given $\hat{K} = k$ and $\mathbf{X}_1 \in A_\varepsilon^{(k)}(X)$, by $P_{col,i}(k)$. Since additional columns in $\tilde{\mathscr{C}}^{(2)}$ would decrease the pairwise collision probability between independent rows, for any $\hat{K} \ge k$, we have

$$P_{col,i}(\hat{K}) \le P_{col,i}(k) \qquad (32)$$

Since we perform exact row matching and the rows of $\tilde{\mathscr{C}}^{(1)}$ are independent, we have

$$P_{col,i}(k) = \Pr(\mathbf{X}_i = \mathbf{X}_1 | \mathbf{X}_1 \in A_\varepsilon^{(k)}(X)) \qquad (33)$$
$$\le 2^{-k(H(X)-\varepsilon)} \qquad (34)$$

Thus, we can bound the probability of error P_e as

$$P_e \le \sum_{i=2}^{2^{nR}} P_{col,i}(k) + \varepsilon + \kappa_n + \mu_n \qquad (35)$$
$$\le 2^{nR}2^{-k(H(X)-\varepsilon)} + \varepsilon + \kappa_n + \mu_n \qquad (36)$$

Since m is exponential in n, by Lemma 1, $\mu_n \to 0$ as $n \to \infty$. Furthermore, $\hat{K}$ is a Binom$(n, 1-\delta)$ random variable and from LLN $\kappa_n \to 0$ as $n \to \infty$. Thus

$$P_e \leq \varepsilon \text{ as } n \to \infty \tag{37}$$

if $R < (1 - \delta - \varepsilon)H(X)$. Thus, we can argue that any database growth rate R satisfying

$$R < (1 - \delta)H(X) \tag{38}$$

is achievable, by taking ε small enough. $\square$

An illustrative example of histogram-based detection (Steps 1-4 of the above proof) is given in Figure 2.

We stress that although we use a rowwise matching scheme after the detection phase rather than matching at the database level, this does not cause a performance loss in terms of achieving the matching capacity, as we prove in Section IV.

IV. CONVERSE

Theorem 1 states that when the repetition pattern is constant across all the rows of $\mathscr{C}^{(2)}$, we can convert repetitions to erasures, achieving the erasure bound. In this section, we show that the lower bound on the matching capacity C given in Section III is, in fact, tight, by proving a tight upper bound on C.

Proof of Converse of Theorem 1. Here we prove that $C = (1 - \delta)H(X)$ is an upper bound on all achievable database growth rates. We adopt a genie-aided proof where the repetition pattern $\mathbf{S}^n$ is available a-priori. Furthermore, we use the modified Fano's inequality presented in [8].

Let P_e be the probability that the scheme is unsuccessful for a given pair of matching rows, R be the database growth rate. Since Θ is a uniform permutation, from Fano's inequality, we have

$$H(\Theta | \mathscr{C}^{(1)}, \mathscr{C}^{(2)}) \leq 1 + P_e \log(m!) \tag{39}$$

$$\frac{1}{m} H(\Theta | \mathscr{C}^{(1)}, \mathscr{C}^{(2)}) \leq \frac{1}{m} + P_e \log m \tag{40}$$

where (40) follows from $m! \leq m^m$. Thus, we get

$$H(\Theta) = H(\Theta | \mathscr{C}^{(1)}, \mathscr{C}^{(2)}) + I(\Theta; \mathscr{C}^{(1)}, \mathscr{C}^{(2)}) \tag{41}$$

$$\frac{1}{m} H(\Theta) \leq \frac{1}{m} + P_e \log m + \frac{1}{m} I(\Theta; \mathscr{C}^{(1)}, \mathscr{C}^{(2)}) \tag{42}$$

$$\frac{1}{mn} H(\Theta) \leq \frac{1}{mn} + \frac{1}{n} P_e \log m + \frac{1}{mn} I(\Theta; \mathscr{C}^{(1)}, \mathscr{C}^{(2)}) \tag{43}$$

and

$$I(\Theta; \mathscr{C}^{(1)}, \mathscr{C}^{(2)}) = I(\Theta; \mathscr{C}^{(2)}) + I(\Theta; \mathscr{C}^{(1)} | \mathscr{C}^{(2)}) \tag{44}$$

$$= I(\Theta; \mathscr{C}^{(1)} | \mathscr{C}^{(2)}) \tag{45}$$

Now, by assuming the availability of the repetition pattern $\mathbf{S}^n$, we can further upper bound the RHS of (45) as

$$I(\Theta; \mathscr{C}^{(1)} | \mathscr{C}^{(2)}) \leq I(\Theta, \mathscr{C}^{(2)}; \mathscr{C}^{(1)}) \tag{46}$$

$$\leq I(\Theta, \mathscr{C}^{(2)}, \mathbf{S}^n; \mathscr{C}^{(1)}) \tag{47}$$

$$= \sum_{i=1}^{m} I(\mathscr{C}_i^{(1)}; \mathscr{C}_{\Theta_i^{-1}}^{(2)}, \mathbf{S}^n) \tag{48}$$

$$= mI(\mathbf{X}^n; \mathbf{Y}, \mathbf{S}^n) \tag{49}$$

where (48)-(49) follow from the fact that the non-matching rows are *i.i.d.* conditioned on the repetition pattern $\mathbf{S}^n$.

Now we prove that the deleted and replicated columns do not offer any additional information. Let $z(\mathbf{s}^n) = \sum_{i=1}^{n} \mathbb{1}_{[s_i=0]}$. Since the entries of $\mathscr{C}^{(1)}$ are *i.i.d.*, we have

$$I(\mathbf{X}^n; \mathbf{Y}, \mathbf{S}^n) = H(\mathbf{X}^n) - H(\mathbf{X}^n | \mathbf{Y}, \mathbf{S}^n) \tag{50}$$

$$= nH(X) - \sum_{\mathbf{s}^n} p(\mathbf{s}^n) H(\mathbf{X}^n | \mathbf{Y}, \mathbf{S}^n = \mathbf{s}^n) \tag{51}$$

$$= nH(X) - \sum_{\mathbf{s}^n} p(\mathbf{s}^n) z(\mathbf{s}^n) H(X) \tag{52}$$

$$= nH(X) - H(X)\mathbb{E}[z(\mathbf{S}^n)] \tag{53}$$

$$= n(1 - \delta)H(X) \tag{54}$$

Thus we have

$$I(\Theta; \mathscr{C}^{(1)} | \mathscr{C}^{(2)}) \leq mn(1 - \delta)H(X) \tag{55}$$

Using Stirling's approximation,

$$\lim_{n \to \infty} \frac{1}{mn} H(\Theta) = \lim_{n \to \infty} \left[\frac{1}{n} \log m - \frac{1}{n} \log e + \frac{1}{mn} O(n) \right] = R \tag{56}$$

Combining (55) and (56), we get

$$\lim_{n \to \infty} \frac{1}{mn} H(\Theta) \leq \lim_{n \to \infty} \left[\frac{1}{mn} + P_e \frac{1}{n} \log m + (1 - \delta)H(X) \right] \tag{57}$$

$$R \leq (1 - \delta)H(X) \tag{58}$$

where (58) follows from the fact that $P_e \to 0$ as $n \to \infty$. $\square$

Equations (50)-(54) suggest that the deleted and the replicated columns do not offer any information. As a result, discarding the deleted columns and the additional replicas in the matching scheme of Section III do not impact optimality. We stress that this is unique to the noiseless setup. As we show in [12], in the noisy setting, while the deleted columns offer no additional information, replicas do help with matching as in a repetition code.

V. CONCLUSION

In this paper, we have studied the database matching problem under random column repetitions. By proving the asymptotic-uniqueness of the column histograms of the databases, we have showed that these histograms can be used for the detection of the deleted and replicated columns. Using the proposed histogram-based detection and exact row matching, we have derived an achievability result, which we have showed is tight, thus giving us the database matching capacity. Our ongoing work includes investigating the matching capacity when different subsets of rows experience different repetition patterns.

REFERENCES

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

[2] J. Sedayao, R. Bhardwaj, and N. Gorade, "Making big data, privacy, and anonymization work together in the enterprise: Experiences and issues," in *2014 IEEE International Congress on Big Data*, 2014, pp. 601–607.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[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] 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.

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

[11] S. Diggavi and M. Grossglauser, "On information transmission over a finite buffer channel," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1226–1237, 2006.

[12] S. Bakirtas and E. Erkip, "Seeded database matching under noisy column repetitions," Available at:
https://serhatbakirtas.github.io/files/seedednoisydbmatching.pdf.

[13] M. Cheraghchi and J. Ribeiro, "An overview of capacity results for synchronization channels," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 3207–3232, 2021.

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