Seeded Database Matching Under Noisy Column Repetitions

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

***Abstract*—Does the authors part look OK? We did it differently last year, but you didn't like that version :)**
Haven't written the abstract. Please skip it. The de-anonymization of user identities from publicly anynomized data by matching various forms of user data available on the internet raises privacy concerns. A fundamental understanding of the privacy leakage in such scenarios requires a careful study of conditions under which correlated databases can be matched. Motivated by synchronization errors in time indexed databases, in this work, matching of random databases under random column deletion is investigated. Adapting tools from information theory, in particular ones developed for the deletion channel, conditions for database matching in the absence and presence of deletion location information are derived, showing that partial deletion information significantly increases the achievable database growth rate for successful matching. Furthermore, given a batch of correctly-matched rows, a deletion detection algorithm that provides partial deletion information is proposed and a lower bound on the algorithm's deletion detection probability in terms of the column size and the batch size is derived. The relationship between the database size and the batch size required to guarantee a given deletion detection probability using the proposed algorithm suggests that a batch size growing double-logarithmic with the row size is sufficient for a nonzero detection probability guarantee.

I. Introduction

The first paragraphs of the intro are the same. I need to rewrite it for the noisy. Please skip until we start describing our work. In the past decades, the collection of personal data has been accelerated, mostly due to the proliferation of smart devices and the emergence of big data applications. The collected personal data, which could be sensitive from a privacy point of view, has been, publicly or found to be, sold or published by companies and/or public institutions, for commercial or academic purposes. Although the publication of the data has been done after *anonymization*, which is the procedure of removing the explicit identifiers, such as names and social security numbers, from the corresponding users, there has been a growing concern over the potential privacy leakage due to the inefficiency of anonymization, from a industrial [1] and legal point of views [2]. In these works, the majority of the arguments regarding the weakness of anonymization on its own have revolved around the potential exploitation of the correlation between the anonymized data and publicly-available data on the users. Indeed, using real data, in [3]–[7], researchers have practically shown that anonymization, by itself, cannot prevent the potentially-menacing privacy leakage. However, although successful, these works have lacked a fundamental and comprehensive understanding of the conditions under which the datasets are prone to privacy attacks.

This work is supported by NYU WIRELESS Industrial Affiliates and National Science Foundation grant CCF-1815821.

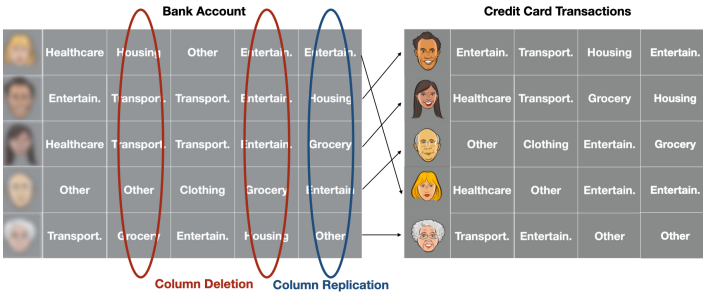

Fig. 1. An illustrative example of database matching under column repetitions. The columns circled in red are deleted whereas the column circled in blue is repeated twice, *i.e.,* replicated.This figure will be replaced.

In the past few years, the de-anonymization of correlated databases have been investigated in [8]–[10] from a fundamental information-theoretic perspective. In [8], Shirani *et al.* assumed a pair of two 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, by using standard information-theoretic arguments. In [9], by assuming a different performance criterion, Cullina *et al.* introduced *cycle mutual information* as a new correlation metric and derived sufficient conditions for a successful matching and a converse result, in terms of this metric. Their probability of error is $\Pr(\hat{\Theta} \neq \Theta)$, whereas ours is $\Pr(\hat{\Theta}(I) \neq \Theta(I))$ for uniform I. Should I mention this in more detail?

In [10], motivated by the synchronization errors in the sampling of time-series datasets, Bakırtaş and Erkip considered database matching under *random column deletions*. Assuming two databases of the same number of users (rows), generated according to a bivariate stochastic process where one of the databases suffers from random column deletions, with unknown deletion indices, similar to the deletion channel model [11]. Importing tools from deletion channel literature [12], assuming a probabilistic side information on the deletion locations and allowing a probability of mismatch vanishing with the number of attributes (columns), they de-

rived an achievable database growth rate. After demonstrating the impact of such side information on achievability, they proposed an algorithm to extract this side information on the deletion locations from a batch of already-matched rows, called *seeds*.

Intro diverges from the noiseless paper from now on.

In a companion paper [**?**], whose shorter version is also submitted to ISIT, we investigate database matching under noiseless column repetition, which is a generalization of work done in [10]. In this model, in addition to some columns being deleted, some of the columns are *replicated*, *i.e.,* sampled multiple times. Assuming the number repetitions of each column is a random integer with an underlying probability distribution, we derive tight sufficient and necessary conditions on the database size, the column repetition distribution, and the database generating distribution for successful matching. We do so, by proving the asymptotic-uniqueness of the types of the columns of the databases and proposing a *type-based detection* scheme which infers the exact repetition pattern from a pair of *unmatched* databases with high probability and a matching scheme, and by utilizing the modified Fano's inequality, proposed in [8].Any mention to the findings, results?

In this paper, our goal is to investigate the sufficient and the necessary conditions for the successful matching of rows under noisy column repetitions, in the presence of noisy seeds. We assume a database model where column repetition is followed by independent noise on the database entries. In this case, the existence of noise prevents us from using the type-based repetition detection algorithm, proposed for the noiseless setup, therefore, we assume seeds, as done in both graph matching ([13], [14]) and database matching ([10]) literatures. Under these assumptions, we devise two detection algorithms, the replication-detection algorithm which does not require seeds and the seeded deletion-detection one. We prove that the asymptotic correctness of the replication detection. Second, we prove that if the seed size B grows in the order of number of columns n, which is logarithmic in the number of rows m of the database, the deletion locations can be extracted from the seeds. Then, we adapt our matching scheme to the noisy setup to derive sufficient for the successful matching. Finally, again, making use of the modified Fano's inequality, we prove a converse result. We show that, as long as $B = \Theta(\log m)$, the sufficient and the necessary conditions are tight up to equality, characterizing the matching capacity of the database matching problem under noisy column repetitions.

The organization of this paper is as follows: Section II contains the formulation of the problem. In Section III, the replication and seeded deletion algorithms, the following matching scheme and the subsequent results on sufficient conditions for the successful database matching are presented. In Section IV, results on necessary conditions for the successful database matching are presented. 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. The logarithms, unless stated explicitly, are in base 2.

Fig. 2. An illustration of the generation process of $\mathscr{C}^{(2)}$.

II. Problem Formulation

We use the following definitions, some of which are taken from [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 discrete support $\mathfrak{X} = \{1,\dots,|\mathfrak{X}|\}$.

Definition 2. (Database Growth Rate, [8]) The *database growth rate* R of an (m,n,p_X) unlabeled database is defined as

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

Definition 3. (Column Repetition Pattern) The *column deletion 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 $\{0,\dots,s_{\max}\}$.

Definition 4. (Labeled Noisy Repeated Database) Let $\mathscr{C}^{(1)}$ be an (m,n,p_X) unlabeled database. Let $\mathbf{S}^n$ be the repetition pattern, Θ be a uniform permutation of $[m]$, independent of $\mathscr{C}^{(1)}$ and $p_{Y|X}$ What should we call $p_{Y|X}$? Channel? DMC? Noise distribution? be a conditional probability distribution with both X and Y taking values from $\mathfrak{X}$. Given $\mathscr{C}^{(1)}$, $\mathbf{S}^n$ and $p_{Y|X}$, the pair $(\mathscr{C}^{(2)}, \Theta)$ is called the *labeled noisy repeated database* if the respective $(i,j)^{\text{th}}$ entries $\mathscr{C}^{(1)}_{i,j}$ and $\mathscr{C}^{(2)}_{i,j}$ of $\mathscr{C}^{(1)}$ and $\mathscr{C}^{(2)}$ have the following relation: Is this a good way to "formalize" the problem?

$$\mathscr{C}^{(2)}_{i,j} = \begin{cases} E, & \text{if } S_j = 0 \\ \mathbf{Y}^{S_j} & \text{if } S_j \geq 1 \end{cases} \quad \forall i \in [m], \forall j \in [n] \tag{2}$$

where $\mathbf{Y}^{S_j}$ is a row random vector of length S_j with the following probability distribution, conditioned on $\mathscr{C}^{(1)}_{\Theta^{-1}(i),j}$.

$$\Pr\left(\mathbf{Y}^{S_j} = \mathbf{y}^{S_j} \middle| \mathscr{C}^{(1)}_{\Theta^{-1}(i),j}\right) = \prod_{l=1}^{S_j} p_{Y|X}\left(y_l \middle| \mathscr{C}^{(1)}_{\Theta^{-1}(i),j}\right) \tag{3}$$

and $\mathscr{C}^{(2)}_{i,j} = E$ corresponds to $\mathscr{C}^{(2)}_{i,j}$ being the empty string.

The relationship between $\mathscr{C}^{(1)}$ and $\mathscr{C}^{(2)}$ is illustrated in Figure 2.

For the labeled and unlabeled databases in Definition 4, the i^{th} 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*.

In this paper we assume that $\mathbf{S}^n$ and $\mathscr{C}^{(1)}$ are independent. Furthermore, Definiton 4 states that S_j indicates the times the j^{th} column of $\mathscr{C}^{(1)}$ is repeated. In the case that $S_j = 0$, the

j^{th} column of $\mathscr{C}$ is said to be *deleted* and when $S_i > 1$, i^{th} column of $\mathscr{C}$ is said to be *replicated*. Furthermore, (3) states that $\mathscr{C}^{(2)}_{i,j}$ is the output of the discrete memoryless channel (DMC) $p_{Y|X}$ with input sequence consisting of S_j copies of $\mathscr{C}^{(1)}_{\Theta^{-1}(i),j}$ concatenated together.

In the noisy setting, inferring the column repetition pattern is a harder task, compared to the noiseless setting, investigated in [**?**], where one could possibly make use of permutation-invariant features of the columns. Therefore, we assume the availability of *seeds*, as done in database matching [10] and graph matching [13], [14] literatures.

Definition 5. (Noisy Seeds) **Seeds are noisy, but should I just call them seeds?** Given the column repetition pattern $\mathbf{S}^n$ and $p_{Y|X}$, a *seed* is a pair of correctly-matched rows $(\mathbf{X},\mathbf{Y})$ with corresponding the repetition pattern $\mathbf{S}^n$, where $\mathbf{Y}$ and $\mathbf{X}$ have the relationship described in Definition 4. Similarly, a *batch of B seeds* is pair of $B \times n$ and $B \times \sum_{j=1}^{n} S_j$ matrices $(\mathscr{D}^{(1)},\mathscr{D}^{(2)})$ where $\mathscr{D}^{(2)}$ is obtained from $\mathscr{D}^{(1)}$, as described in Definition 4, with a known identity labeling function. In our work, we assume that the *seed size* B is a function of n.

Definition 6. (Successful Matching Scheme) Given a batch of B seeds, a *matching scheme* is a sequence of mappings $\phi_n : (\mathscr{C}^{(1)},\mathscr{C}^{(2)}) \to \hat{\Theta}_n$ where $\hat{\Theta}_n \in [m]^m$ is the estimate of the correct labeling function Θ_n. The scheme ϕ_n is *successful* if

$$\Pr\left(\Theta_n(I) = \hat{\Theta}_n(I)\right) \to 1 \text{ as } n \to \infty \tag{4}$$

where the index I is drawn uniformly from $[m]$. Here, the dependence of $\hat{\Theta}_n$ on the batch of seeds is omitted for brevity.

Definition 7. (Achievable Database Growth Rate) Given a database probability distribution p_X, a repetition probability distribution p_S, $p_{Y|X}$ and a batch of B seeds, 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 8. (Matching Capacity) Given a database probability distribution p_X, a repetition probability distribution p_S, $p_{Y|X}$ and a batch of B seeds, the *matching capacity* $C(B)$ is the supremum of the set of all achievable rates.

III. Main Result and Achievability

In this section, we present our main result, stated in the following theorem and prove its achievability, by generalizing the detection and matching schemes devised for the noiseless setting in [**?**], developing new ones when necessary. We present the proof of the converse in Section IV.

Theorem 1. ***(Matching Capacity)*** *Given a database probability distribution p_X, a repetition probability distribution p_S, $p_{Y|X}$ and a batch of B seeds, the matching capacity is*

$$C(B) = I(X;\mathbf{Y}(X,S),S) \tag{5}$$

if $B = \omega(n)$, where $\mathbf{Y}(X,S) = Y_1,\ldots,Y_S$, Y_i being noisy copies of X, independent of each other conditioned on X.

Before discussing the proof of Theorem 1, we present the following corollary.

Corollary 1. ***(Noiseless Setting)*** *In the noiseless setting, where*

$$p_{Y|X}(y|x) = \mathbb{1}_{[y=x]}\, \forall x \in \mathfrak{X} \tag{6}$$

our main result becomes

$$C(B) = (1-\delta)H(X) \tag{7}$$

where $\delta \triangleq p_S(0)$ is the deletion probability.

In fact, as we discuss in Section III-A, in the noiseless setting, this result applies to any seed size B.

To prove the achievability, we consider a three phase matching strategy, described in the following subsections. In Section III-A, we discuss our noisy replica detection algorithm and prove an asymptotic performance result. Then, in Section III-B, we discuss the extraction of deletion locations from the seeds and derive the seed size sufficient for an asymptotic performance guarantee. Finally, in Section III-C, we prove the achievability by generalizing the rowwise matching scheme, proposed in [**?**], for the noiseless scenario.

Throughout this section, due to space constraints, we make the following assumptions on the distribution $p_{Y|X}$.

Assumption 1.

$$\sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} p_X(x)p_{Y|X}(y|x)\left[1-p_Y(y)\right] > \sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} p_X(x)p_{Y|X}(y|x)\left[1-p_{Y|X}(y|x)\right] \tag{8}$$

Assumption 2.

$$\sum_{x_1\in\mathfrak{X}}\sum_{x_2\in\mathfrak{X}} p_X(x_1)p_X(x_2)[1-p_{Y|X}(x_2|x_1)] > \sum_{x\in\mathfrak{X}} p_X(x)[1-p_{Y|X}(x|x)] \tag{9}$$

Assumption 1 implies that the two noisy entries originated from two independent entries are more likely to be different than two noisy replicas of a single entry. On the other hand, Assumption 2 requires that a noisy entry is more likely to be equal to the entry it is originated from, compared to an independent noiseless entry. In fact, these assumptions can either be automatically satisfied or circumvented by simply assuming $p_{Y|X} \neq p_Y$, *i.e.,* a positive correlation between the databases, which is a rational assumption. Indeed, we prove these in our journal paper. Do we need to further mention that we rigorously proved them?

A. Noisy Replica Detection

Given $\mathscr{C}^{(1)}$ and $\mathscr{C}^{(2)}$, we detect the replications, by extracting permutation-invariant features of the columns, similar to [**?**]. In [**?**], we choose the type of each column as its permutation-invariant feature and prove that these types are asymptotically unique. Then we match the types of the columns of both databases to infer the repetition process. In the noisy setup, although still asymptotically-unique, the types

of the columns of the two databases could not be matched in the same way. Therefore, we propose a replica detection algorithm which adopts the Hamming distance between the columns of $\mathscr{C}^{(2)}$ as the permutation-invariant feature.

Let $\mathbf{R}_i$ denote the i^{th} column of $\mathscr{C}^{(2)}$, K denote the number of columns of $\mathscr{C}^{(2)}$. Our replica detection algorithm works as follows: We first compute the Hamming distances $d_H(\mathbf{R}_i,\mathbf{R}_{i+1})$, for $i\in[K-1]$ between $\mathbf{R}_i$ and $\mathbf{R}_{i+1}$. If $d_H(\mathbf{R}_i,\mathbf{R}_{i+1})>m\tau$, where the test threshold τ will be specified in Lemma 2, the algorithm concludes that $\mathbf{R}_i$ and $\mathbf{R}_{i+1}$ are independent. Otherwise, it concludes that $\mathbf{R}_i$ and $\mathbf{R}_{i+1}$ are two noisy replicas of a single column. We show that this algorithm can infer the replications with high probability, in the following lemma, proven in [**?**].

Lemma 2. ***(Noisy Replica Detection)*** *Suppose that $p_{Y|X}$ satisfies Assumption 1. Let ε_i denote the event that the algorithm, described above, deduces that $\mathbf{R}_i$ and $\mathbf{R}_{i+1}$ are noisy replicas, for $i\in[K-1]$ and*

$$p_0=\sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{X}}p_X(x)p_{Y|X}(y|x)\left[1-p_Y(y)\right] \tag{10}$$

$$p_1=\sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{X}}p_X(x)p_{Y|X}(y|x)\left[1-p_{Y|X}(y|x)\right] \tag{11}$$

Then, for any threshold τ bounded away from p_0 and p_1,

$$\Pr\left(\varepsilon_i=\mathbb{1}_{[\mathbf{R}_i,\mathbf{R}_{i+1}\text{ are noisy replicas.}]}\,\forall i\right)\to 1 \text{ as } n\to\infty \tag{12}$$

Should I give a sketch?

B. Seeded Deletion Detection

Let $(\mathscr{D}^{(1)},\mathscr{D}^{(2)})$ be a batch of B seeds, undergone the same repetition pattern $\mathbf{S}^n$ as $(\mathscr{C}^{(1)},\mathscr{C}^{(2)})$. Our deletion detection algorithm works as follows: First, we perform replica detection via the replica detection algorithm discussed in Section III-A. After finding the replicas, we discard all-but-one of the noisy replicas from $\mathscr{D}^{(2)}$, to obtain $\tilde{\mathscr{D}}^{(2)}$, whose column size is denoted by $\tilde{K}$. At this step, the detection problem is reduced into a deletion-only one. Then, we perform an exhaustive search over all potential deletion patterns and outputs the one minimizing total Hamming distance with $\tilde{\mathscr{D}}^{(2)}$, denoted by $\hat{I}_{\text{del}}$. In other words,

$$\hat{I}_{\text{del}}=\underset{I\subseteq[n],|I|=n-\tilde{K}}{\operatorname{argmin}}\; d_H(\mathscr{D}^{(1)}([n]\setminus I),\tilde{\mathscr{D}}^{(2)}) \tag{13}$$

where $\mathscr{D}^{(1)}([n]\setminus I)$ denotes the matrix obtained by discarding the columns whose indices lie in I. We show that if the seed size $B=\omega(n)$, this algorithm can infer the deletion locations with high probability, in the following lemma, proven in [**?**].

Lemma 3. ***(Noisy Seeded Deletion Detection)*** *Suppose that $p_{Y|X}$ satisfies Assumption 1 and Assumption 2. Given the repetition pattern $\mathbf{S}^n$, let $I_{del}=\{j\in[n]|S_j=0\}$. If $B=\omega(n)$,*

$$\Pr\left(\hat{I}_{del}=I_{del}\right)\to 1 \text{ as } n\to\infty \tag{14}$$

Again, is a sketch required?

After we perform the replica and deletion detection, as discussed above, we obtain an estimate $\hat{\mathbf{S}}^n$ of the repetition pattern $\mathbf{S}^n$.

C. Matching Scheme

After discussing the construction of $\hat{\mathbf{S}}^n$, we are ready to discuss the proof of the achievability, stated in the following theorem.

Theorem 4. ***(Achievability Result)*** *Consider an (m,n,p_X) unlabeled database, a column repetition pattern $\mathbf{S}^n$ with repetition distribution p_S, $p_{Y|X}$ satisfying Assumption 1 and Assumption 2 and a seed size $B=\omega(n)$. Then, any database growth rate $R>0$ is achievable if*

$$R<I(X;\mathbf{Y}(X,S),S) \tag{15}$$

Proof (Sketch). Let $\mathbf{S}^n$ be the underlying repetition pattern. Now, we perform the following steps:

- **Step 1.** Perform replica detection as described in Section III-A, whose probability of error is denoted by ρ_n.
- **Step 2.** Perform seeded deletion detection as described in Section III-B, whose probability of error is denoted by μ_n.
- **Step 3.** Discard the deleted columns from $\mathscr{C}^{(1)}$, to obtain $\tilde{\mathscr{C}}^{(1)}$, whose column size denoted by $\tilde{K}=n-\sum_{i=1}^{n}\mathbb{1}_{[S_i=0]}$.
- **Step 4.** Place markers between the noisy replica runs of different columns to obtain $\tilde{\mathscr{C}}^{(2)}$. Note that provided that the detection algorithms performed correctly, there are exactly $\tilde{K}$ such runs, where the j^{th} run in $\tilde{\mathscr{C}}^{(2)}$ corresponds to the noisy copies of the j^{th} column of $\Theta\circ\tilde{\mathscr{C}}^{(1)}$. An illustrative example perhaps?
- **Step 5.** Fix $\varepsilon>0$. If $\tilde{K}<k=n(1-\delta-\varepsilon)$, declare error, whose probability is denoted by κ_n. Otherwise, proceed with the next step.
- **Step 6.** Match the l^{th} row $\mathbf{Y}_l$ of $\tilde{\mathscr{C}}^{(2)}$ with the i^{th} row $\mathbf{X}_i$ of $\tilde{\mathscr{C}}^{(2)}$, if $\mathbf{X}_i$ is the only row of $\tilde{\mathscr{C}}^{(1)}$, jointly ε-typical with $\mathbf{Y}_l$, assigning $\hat{\Theta}(i)=l$. Otherwise, declare error.

After following the steps given above, we can use standard information-theoretic arguments to bound the probability of error of this scheme as follows

$$P_e\le 2^{nR}2^{-n(I(X;\mathbf{Y}(X,S),S)-3\varepsilon)}+\varepsilon+\rho_n+\mu_n+\kappa_n \tag{16}$$

Note that since m is exponential in n, $B=\omega(n)$, and from WLLN, we have $\rho_n\to 0$, $\mu_n\to 0$ and $\kappa_n\to 0$ as $n\to\infty$. Thus,

$$P_e\le\varepsilon \text{ as } n\to\infty \tag{17}$$

if $R<I(X;\mathbf{Y}(X,S),S)$, concluding the proof. □

The matching scheme proposed above, for the noisy database matching diverges from the one proposed in [**?**] for the noiseless setting in the following way: First, in the noiseless setting, the seeds are not required and one detection algorithm is enough to deduce the deletions and replications. Second, in Step 4 above, we retained the noisy replicas whereas we discard the noiseless replicas in [**?**]. This is because after replica detection, noisy replicas offer additional information, unlike the noiseless setting. In this case, we basically have a repetition code of the columns of

the unlabeled database, with varying length. Finally, since we retain the replicas, the rows $\mathbf{Y}_l$ and $\mathbf{X}_i$ described in Step 6 have different lengths. Therefore we switch to a joint typicality based matching of rows instead of the one based on exact sequence matching, as proposed in [**?**].

The repetition code argument given above leads the following distinction between the noiseless database matching, the noisy database matching, and channel decoding: The existence of replicas in the noiseless setting does not have any effect on the matching capacity. On the other hand, the replicas are helpful in the noisy setting. Finally, the replications make the matching difficult in channel decoding, decreasing the corresponding channel capacity.

IV. The Converse Result

Theorem 4 states that when the repetition pattern is constant across all the rows of $\mathscr{C}^{(1)}$, under some mild assumptions on the noise $p_{Y|X}$, we can detect replicas automatically and deleted columns through a seed size growing faster than the column size. In the following theorem, we show that the database growth rate achieved with this seed size, is in fact an upper bound on all achievable rates for any seed size.

Theorem 5. ***(Converse Result)*** *Consider an (m,n,p_X) unlabeled database, a column repetition pattern $\mathbf{S}^n$ with repetition distribution p_S and a DMC $p_{Y|X}$. Then, for any seed size B, a database growth rate R is achievable only when*

$$R \leq I(X;\mathbf{Y}(X,S),S) \tag{18}$$

Proof. We prove this theorem using the modified Fano's inequality presented in [8]. Let P_e be the probability that the scheme is unsuccessful for a pair of matching rows. Since Θ is a uniform permutation, from [8], we have

$$\frac{1}{mn}H(\Theta) \leq \frac{1}{mn} + \frac{1}{n}P_e \log m + \frac{1}{mn} I(\Theta;\mathscr{C}^{(1)},\mathscr{C}^{(2)},\mathscr{D}^{(1)},\mathscr{D}^{(2)}) \tag{19}$$

$$I(\Theta;\mathscr{C}^{(1)},\mathscr{C}^{(2)},\mathscr{D}^{(1)},\mathscr{D}^{(2)}) = I(\Theta;\mathscr{C}^{(1)}|\mathscr{C}^{(2)},\mathscr{D}^{(1)},\mathscr{D}^{(2)}) \tag{20}$$

Furthermore, we have

$$\begin{aligned} I(\Theta;\mathscr{C}^{(1)},\mathscr{C}^{(2)},\mathscr{D}^{(1)},\mathscr{D}^{(2)}) &= I(\Theta;\mathscr{C}^{(1)},\mathscr{C}^{(2)}|\mathscr{D}^{(1)},\mathscr{D}^{(2)}) \\ &\quad + I(\Theta;\mathscr{D}^{(1)},\mathscr{D}^{(2)}) && (21)\\ &= I(\Theta;\mathscr{C}^{(1)},\mathscr{C}^{(2)}|\mathscr{D}^{(1)},\mathscr{D}^{(2)}) && (22)\end{aligned}$$

From the independence of Θ, $\mathscr{C}^{(2)}$ and $(\mathscr{D}^{(1)},\mathscr{D}^{(2)})$, we get

$$\begin{aligned} I(\Theta;\mathscr{C}^{(1)},\mathscr{C}^{(2)}|\mathscr{D}^{(1)},\mathscr{D}^{(2)}) &= I(\Theta;\mathscr{C}^{(1)}|\mathscr{C}^{(2)},\mathscr{D}^{(1)},\mathscr{D}^{(2)}) \\ &\quad + I(\Theta;\mathscr{C}^{(2)}|\mathscr{D}^{(1)},\mathscr{D}^{(2)}) && (23)\\ &= I(\Theta;\mathscr{C}^{(1)}|\mathscr{C}^{(2)},\mathscr{D}^{(1)},\mathscr{D}^{(2)}) && (24)\end{aligned}$$

Now we can further upper bound the RHS of (24) as

$$\begin{aligned} I(\Theta;\mathscr{C}^{(1)}|\mathscr{C}^{(2)},\mathscr{D}^{(1)},\mathscr{D}^{(2)}) &\leq I(\Theta,\mathscr{C}^{(2)},\mathscr{D}^{(1)},\mathscr{D}^{(2)};\mathscr{C}^{(1)}) && (25)\\ &\leq I(\Theta,\mathscr{C}^{(2)},\mathbf{S}^n;\mathscr{C}^{(1)}) && (26)\\ &\leq \sum_{i=1}^{m} I(\mathscr{C}_i^{(1)};\mathscr{C}^{(2)}_{\Theta_i^{-1}},\mathbf{S}^n) && (27)\\ &= mI(\mathbf{X}^n;\mathbf{Y},\mathbf{S}^n) && (28)\\ &= mnI(X;\mathbf{Y}(X,S),S) && (29)\end{aligned}$$

where (26) follows from the fact that given $\mathbf{S}^n$, $\mathscr{D}^{(1)},\mathscr{D}^{(2)}$ do not offer any additional information. (27) follows from the independence of the non-matching rows conditioned on the repetition pattern $\mathbf{S}^n$ and (28) holds because the rows of $\mathscr{C}^{(1)}$ and the noise on the entries of $\mathscr{C}^{(2)}$ are *i.i.d.*. Finally (29) follows from the database entries being *i.i.d.*.

We stress that, since we assumed the availability of $\mathbf{S}^n$ in (26), this bound holds for any seed size B, as seeds can only contribute to deletion and replica detection.

So far, we have

$$\frac{1}{m}H(\Theta) \leq \frac{1}{m} + P_e \log m + nI(X;\mathbf{Y}(X,S),S) \tag{30}$$

We now look at $\lim_{n\to\infty}\frac{1}{mn}H(\Theta)$, using Stirling's approximation.

$$H(\Theta) = \log m! = m\log m - m\log e + O(\log m) \tag{31}$$

$$\lim_{n\to\infty}\frac{1}{mn}H(\Theta) = R \tag{32}$$

Using (32), we get

$$\begin{aligned} \frac{1}{mn}H(\Theta) &\leq \frac{1}{mn} + P_e\frac{1}{n}\log m + I(X;\mathbf{Y}(X,S),S) && (33)\\ \lim_{n\to\infty}\frac{1}{mn}H(\Theta) &\leq \lim_{n\to\infty}\left[\frac{1}{mn} + P_eR + I(X;\mathbf{Y}(X,S),S)\right] && (34)\\ R &\leq I(X;\mathbf{Y}(X,S),S) && (35)\end{aligned}$$

where (35) follows from the fact that $P_e \to 0$ as $n \to \infty$. □

V. Conclusion

In this work, we have studied the database matching problem under random noisy column repetitions. By exploiting the identicality of the repetition pattern across different rows, we have showed that the running Hamming distance between the columns of the labeled noisy repeated database can be used as a permutation-invariant feature of the columns which can be used for the detection of the replications. Furthermore, we have showed that by assuming the availability of seeds, we can infer the locations of the deleted columns through an exhaustive search over the set of possible deletion indices, provided that the seed size grows faster than the column size, which is logarithmic with the original database row size. Using the proposed detection algorithms, and a joint typicality based rowwise matching scheme, we have derived an achievable database growth rate. Then, using a modified version of Fano's inequality, we have showed that this achievability result is indeed tight up to equality. Finally, we show that our matching capacity boils down to the noiseless matching capacity we derive in [**?**], when the noise is removed form the system.

REFERENCES

[1] 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.

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

[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 *2008 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. Houmansadrand, D. 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, 2018.

[8] F. Shirani, S. Garg, and E. Erkip, "A concentration of measure approach to database de-anonymization," in *2019 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 *2018 IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 651–655.

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

[11] M. Cheraghchi and J. Ribeiro, "An overview of capacity results for synchronization channels," *IEEE Trans. Inf. Theory*, 2020.

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

[13] F. Shirani, S. Garg, and E. E., "Seeded graph matching: Efficient algorithms and theoretical guarantees," in *2017 51st Asilomar Conference on Signals, Systems, and Computers*, 2017, pp. 253–257.

[14] D. Fishkind, S. Adali, H. Patsolic, L. Meng, D. Singh, V. Lyzinski, and C. Priebe, "Seeded graph matching," *Pattern Recognition*, vol. 87, pp. 203–215, 2019.