Seeded Database Matching Under Noisy Column Repetitions

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

***Abstract*—The re-identification or de-anonymization of users from anonymized data through matching with publicly-available correlated user data has raised privacy concerns, leading to the complementary measure of obfuscation in addition to anonymization. Recent research provides a fundamental understanding of the conditions under which privacy attacks are successful, either in the presence of obfuscation or synchronization errors stemming from the sampling of time-indexed databases. This paper presents a unified framework considering both obfuscation and synchronization errors and investigates the matching of databases under noisy column repetitions. By devising replica detection and seeded deletion detection algorithms, and using information-theoretic tools, sufficient conditions for successful matching are derived. It is shown that a seed size logarithmic in the row size is enough to guarantee the detection of all deleted columns. It is also proved that this sufficient condition is necessary, thus characterizing the database matching capacity of database matching under noisy column repetitions and providing insights on privacy-preserving publication of anonymized and obfuscated time-indexed data.**

I. Introduction

With the exponential boom in smart devices and the growing popularity of big data, companies and institutions have been gathering more and more personal data from users which is then either published or sold for research or commercial purposes. Although the published data is typically *anonymized*, *i.e.,* explicit identifiers of the users, such as names and dates of birth are removed, researchers [1] and companies [2] have articulated their concerns over the insufficiency of anonymization for privacy as demonstrated by a series of practical attacks on real data [3]–[7]. *Obfuscation*, which refers to the deliberate addition of noise to the database entries, has been suggested as an additional measure to protect privacy [6]. While extremely valuable, this line of work does not provide a fundamental and rigorous understanding of the conditions under which anonymized and obfuscated databases are prone to privacy attacks.

Recently, matching correlated pairs of databases have been investigated from an information-theoretic [8]–[12] and statistical [13] points of view. In [8], Cullina *et al.* proposed *cycle mutual information* as a metric of correlation and derived sufficient and necessary conditions for successful matching, with the performance criterion being the perfect recovery for all users. In [9], Shirani *et al.* considered a pair of anonymized and obfuscated databases and drew analogies between database matching and channel decoding. By doing so, they derived sufficient and necessary conditions on the *database growth rate* for reliable matching, in the presence of noise on the database entries. In [10] Dai *et al.* investigated the matching of correlated databases with Gaussian attributes with the perfect recovery criterion. In [13], Kunisky and Niles-Weed investigated the same problem as Dai *et al.*, from a statistical perspective, in different database size regimes for several performance criteria.

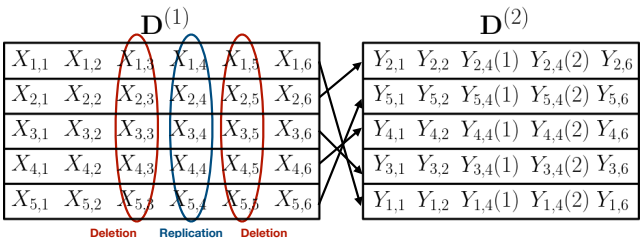

Fig. 1. An illustrative example of database matching under noisy column repetitions. The columns circled in red are deleted whereas the fourth column, which is circled in blue, is repeated twice, *i.e.,* replicated. For each (i,j), $Y_{i,j}$ is the noisy observation of $X_{i,j}$. Furthermore, for each i, $Y_{i,4}(1)$ and $Y_{i,4}(2)$ are noisy replicas of $X_{i,4}$. Our goal is to estimate the row permutation Θ_n which is in this example given as; $\Theta_n(1)=5$, $\Theta_n(2)=1$, $\Theta_n(3)=4$, $\Theta_n(4)=3$ and $\Theta_n(5)=2$, by matching the rows of $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$. Here the i^{th} row of $\mathbf{D}^{(1)}$ corresponds to the $\Theta_n(i)^{\text{th}}$ row of $\mathbf{D}^{(2)}$.

In [11], motivated by the synchronization errors in the sampling of time-series datasets, we investigated the matching of two databases of the same number of users (rows), but with different numbers of attributes (columns). In our model, one of the databases suffers from *random column deletions*, where the deletion indices are only partially and probabilistically available at the matching side. Under this side information assumption, we derived an achievable database growth rate. Demonstrating the impact of this side information on the achievable rate, we then proposed a *deletion detection* algorithm given a batch of correctly-matched rows, *i.e., seeds* and derived the seed size sufficient to guarantee a non-zero deletion detection probability.

In [12], we investigated the matching of Markov databases, thus modeling correlations of the attributes (columns) under noiseless random column repetitions, a non-trivial extension of [11], where the attributes were assumed *i.i.d.*. Under this generalized model, we devised a *column histogram-based* repetition detection algorithm and derived an improved achievable

This work is supported by National Science Foundation grants 1815821 and 2148293.

rate, which is equal to the erasure bound [14]. We then proved a converse showing the tightness of this achievable rate, thereby characterizing the exact matching capacity of Markov database matching under noiseless column repetitions.

In this paper, our goal is to investigate the necessary and the sufficient conditions for the successful matching of database rows under *noisy* column repetitions. We assume a generalized database model where synchronization errors, in the form of column repetitions, are followed by noise, in the form of independent noise on the database entries, as illustrated in Figure 1. The presence of noise prevents us from using the column histogram-based repetition detection algorithm of [12] and unlike [12] requires *seed* users whose identities are known in both databases [11], [15], [16]. Under these assumptions, we devise two algorithms: one for deletion detection and the other for replica detection. We show that if the seed size B grows linearly with the number of columns n, which is assumed to be logarithmic in the number of rows m_n of the database, deletion locations can be extracted from the seeds. Then, we propose a joint typicality-based row matching scheme to derive sufficient conditions for successful matching. Finally, we prove a tight converse result, 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, our main result on the matching capacity and its proof are presented. Finally, in Section IV the results and ongoing work are discussed.

Notation: We denote the set of integers $\{1,...,n\}$ as $[n]$, and matrices with uppercase bold letters. For a matrix $\mathbf{D}$, $D_{i,j}$ denotes the $(i,j)^{\text{th}}$ entry. Furthermore, by A^n, we denote a row vector consisting of scalars $A_1,\ldots,A_n$ and the indicator of event E by $\mathbb{1}_E$. The logarithms, unless stated explicitly, are in base 2. When the distinction is clear from the context, we use Θ to denote either the labeling function or the big theta notation for the asymptotic behavior.

II. Problem Formulation

We use the following definitions, some of which are similar to [9], [11], [12], to formally describe our problem.

Definition 1. (Unlabeled Database) An (m_n,n,p_X) *unlabeled database* is a randomly generated $m_n \times n$ matrix $\mathbf{D}=\{D_{i,j}\in\mathfrak{X}\}$ with *i.i.d.* entries drawn according to the distribution p_X with a finite discrete support $\mathfrak{X}=\{1,\ldots,|\mathfrak{X}|\}$.

Definition 2. (Column Repetition Pattern) The *column repetition pattern* $S^n=\{S_1,S_2,...,S_n\}$ is a random vector consisting of n *i.i.d.* entries drawn from a discrete probability distribution p_S with a finite integer support $\{0,\ldots,s_{\max}\}$.

Definition 3. (Labeled Noisy Repeated Database) Let $\mathbf{D}^{(1)}$ be an (m_n,n,p_X) unlabeled database. Let S^n be the independent repetition pattern, Θ_n be a uniform permutation of $[m_n]$, independent of $(\mathbf{D}^{(1)},S^n)$ and $p_{Y|X}$ be a conditional probability distribution with both X and Y taking values from $\mathfrak{X}$. Given $\mathbf{D}^{(1)}$, S^n and $p_{Y|X}$, $\mathbf{D}^{(2)}$ is called the *labeled noisy repeated database* if the respective $(i,j)^{\text{th}}$ entries $D^{(1)}_{i,j}$ and $D^{(2)}_{i,j}$ of $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$ have the following relation:

$$D^{(2)}_{i,j}=\begin{cases}E, & \text{if } S_j=0\\ Y_i^{S_j} & \text{if } S_j\geq 1\end{cases}\qquad \forall i\in[m_n],\forall j\in[n] \tag{1}$$

where $Y_i^{S_j}$ is a random row vector of length S_j with the following probability distribution, conditioned on $D^{(1)}_{\Theta_n^{-1}(i),j}$

$$\Pr\left(Y_i^{S_j}=y^{S_j}\middle|D^{(1)}_{\Theta_n^{-1}(i),j}\right)=\prod_{l=1}^{S_j}p_{Y|X}\left(y_l\middle|D^{(1)}_{\Theta_n^{-1}(i),j}\right) \tag{2}$$

where $y^{S_j}=y_1,\ldots,y_{S_j}$ and $D^{(2)}_{i,j}=E$ corresponds to $D^{(2)}_{i,j}$ being the empty string.

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

The i^{th} row of $\mathbf{D}^{(2)}$ is said to correspond to the $\Theta_n^{-1}(i)^{\text{th}}$ row of $\mathbf{D}^{(1)}$, where Θ_n is called the *labeling function*.

The relationship between $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$, as described in Definiton 3, is illustrated in Figure 2.

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

Note that (2) states that we can treat $D^{(2)}_{i,j}$ as the output of the discrete memoryless channel (DMC) $p_{Y|X}$ with input sequence consisting of S_j copies of $\mathbf{D}^{(1)}_{\Theta_n^{-1}(i),j}$ concatenated together. We stress that $p_{Y|X}$ is a general model, capturing any distortion and noise on the database entries, though we only refer to this as "noise" in this paper.

As we will discuss in Section III-A, in the noisy setting, inferring the column repetition pattern, particularly deletions, is a harder task compared to the noiseless setting investigated in [12]. Therefore, we assume the availability of *seeds*, as done in noiseless database matching [11] and graph matching [15], [16] literatures.

Definition 4. (Seeds) For the unlabeled and labeled databases in Definitions 1 and 3, a *seed* is a pair of correctly-matched rows. A *batch of B seeds* $(\mathbf{G}^{(1)},\mathbf{G}^{(2)})$ is a pair of databases (sub-matrices) with respective sizes $B\times n$ and $B\times\sum_{j=1}^n S_j$. We assume a polynomial *seed size* $B=\Theta(n^d)$ where d is called the *seed order*.

Definition 5. (Successful Matching Scheme) A *matching scheme* is a sequence of mappings $\phi_n:(\mathbf{D}^{(1)},\mathbf{D}^{(2)},\mathbf{G}^{(1)},\mathbf{G}^{(2)})\mapsto\hat{\Theta}_n$ where $\mathbf{D}^{(1)}$ is the unlabeled database, $\mathbf{D}^{(2)}$ is the labeled noisy repeated database, $(\mathbf{G}^{(1)},\mathbf{G}^{(2)})$ are seeds and $\hat{\Theta}_n$ is the estimate of the correct labeling function Θ_n. The scheme ϕ_n is *successful* if

$$\Pr\left(\Theta_n(J)\neq\hat{\Theta}_n(J)\right)\to 0 \text{ as } n\to\infty \tag{3}$$

where the index J is drawn uniformly from $[m_n]$.

Note that for a given column size n, as the row size m_n increases, so does the probability of mismatch, as a result of having a larger number of candidates. Thus, in order to characterize the relationship between m_n and n, we use the *database growth rate* introduced in [9]. As stated in [13, Theorem 1.2], for distributions with parameters constant in n, the regime of interest is the logarithmic regime where $n \sim \log m_n$.

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

$$R = \lim_{n\to\infty} \frac{1}{n} \log m_n. \tag{4}$$

Definition 7. (Achievable Database Growth Rate) Consider a sequence of (m_n, n, p_X) unlabeled databases, a repetition probability distribution p_S, a noise distribution $p_{Y|X}$ and the resulting sequence of labeled noisy repeated databases. For a seed order d, a database growth rate R is said to be *achievable* if there exists a successful matching scheme when the unlabeled database has growth rate R.

Definition 8. (Matching Capacity) The *matching capacity* $C(d)$ is the supremum of the set of all achievable rates corresponding to a database distribution p_X, a repetition probability distribution p_S, a noise distribution $p_{Y|X}$ and a seed order d.

In this paper, our goal is to characterize the matching capacity $C(d)$, by providing database matching schemes as well as a tight upper bound on all achievable database growth rates.

III. Main Result

In this section, we present our main result on the matching capacity under noisy column repetitions (Theorem 1) and prove its achievability by proposing a three-step approach: *i)* noisy replica detection and *ii)* deletion detection using seeds, followed by *iii)* a row matching algorithm. Then, we outline the proof of the converse.

Theorem 1. ***(Matching Capacity Under Noisy Column Repetitions)*** *Consider a database distribution p_X, a column repetition distribution p_S and a noise distribution $p_{Y|X}$. Then, for any seed order $d \geq 1$, the matching capacity is*

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

where $S \sim p_S$ and $Y^S = Y_1, \ldots, Y_S$ such that

$$\Pr(Y^S = y_1, \ldots, y_S | X = x) = \prod_{i=1}^{S} p_{Y|X}(y_i|x) \tag{6}$$

Theorem 1 states that although the repetition pattern S^n is not known a-priori, given a seed order $d \geq 1$, we can achieve a database growth rate as if we knew S^n. Since the utility of seeds increase with the seed order d, we will focus on $d = 1$, which we show is sufficient to achieve the matching capacity. As we discuss in Section III-D, the converse result holds for any seed size, whereas a general achievability result for the noisy case with $d < 1$ requires additional combinatorial arguments and is omitted due to the space constraints.

Remark 1. (**Noiseless Setting**) Using [12, Corollary 1], we can argue that in the noiseless setting, where

$$p_{Y|X}(y|x) = \mathbb{1}_{[y=x]} \, \forall x \in \mathfrak{X} \tag{7}$$

we have

$$C(d) = (1-\delta) H(X) \tag{8}$$

for any seed order d, where $\delta \triangleq p_S(0)$ is the deletion probability. Furthermore, we show in [12] that in the noiseless setting $Y^S = X \otimes 1^S$, the replicas do not offer any additional information. Thus, for any seed order $d \geq 1$, Theorem 1 agrees with [12, Corollary 1] in the noiseless setting with *i.i.d.* columns.

Remark 2. (**No Synchronization Errors**) As discussed in [9, Corollary 1], when there are no synchronization errors, *i.e.,* $p_S(1) = 1$, we have

$$C(d) = I(X;Y) \tag{9}$$

for any seed order d. Thus, under no synchronization errors, for any seed order $d \geq 1$, Theorem 1 agrees with [9, Corollary 1].

The rest of this section is on the proof of Theorem 1. In Section III-A, we discuss our noisy replica detection algorithm and prove its asymptotic performance. In Section III-B, we introduce a deletion detection algorithm which uses seeds and derive a seed size sufficient for an asymptotic performance guarantee. Then, in Section III-C, we combine these two algorithms and prove the achievability of Theorem 1 by generalizing the rowwise matching scheme proposed in [12] to the noisy scenario. Finally, in Section III-D we present the outline of the proof of the converse of Theorem 1.

Note that when the two databases are independent, Theorem 1 states that the matching capacity becomes zero, hence our results trivially hold. Hence throughout this section, we assume that the two databases are not independent.

A. Noisy Replica Detection

We propose to detect the replicas by extracting permutation-invariant features of the columns of $\mathbf{D}^{(2)}$. Our algorithm only considers the columns of $\mathbf{D}^{(2)}$ and as such, can only detect replications, not deletions. Furthermore, we stress that our replica detection algorithm does not require any seeds.

In [12], we chose the histogram of each column as its permutation-invariant feature, proved that the asymptotic uniqueness of the histograms and matched the column histograms of $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$ to infer the repetition pattern. In the noisy setup, although still asymptotically-unique, the column histograms of the two databases cannot be matched due to noise. Joint typicality arguments do not work either, since arbitrary pairs of column histograms are likely to be jointly typical, even though the columns are independent. Therefore, we propose a replica detection algorithm which only considers

$\mathbf{D}^{(2)}$ and adopts the *Hamming distance between consecutive columns* of $\mathbf{D}^{(2)}$ as the permutation-invariant feature.

Let K denote the number of columns of $\mathbf{D}^{(2)}$, $C_j^{m_n}$ denote the j^{th} column of $\mathbf{D}^{(2)}$, $j = 1,\ldots,K$. Our replica detection algorithm works as follows: We first compute the Hamming distances $d_H(C_j^{m_n}, C_{j+1}^{m_n})$ between $C_j^{m_n}$ and $C_{j+1}^{m_n}$, for $j \in [K-1]$. For some average Hamming distance threshold τ chosen based on $p_{X,Y}$, the algorithm decides that $C_j^{m_n}$ and $C_{j+1}^{m_n}$ are replicas only if $d_H(C_j^{m_n}, C_{j+1}^{m_n}) < m_n\tau$, and independent otherwise. In the following lemma, we show that this algorithm can infer the replicas with high probability.

Lemma 1. ***(Noisy Replica Detection)*** *Let E_j denote the event that the Hamming distance based algorithm described above fails to infer the correct relationship between the columns $C_j^{m_n}$ and $C_{j+1}^{m_n}$ of $\mathbf{D}^{(2)}$, $j = 1,\ldots,K-1$. Then*

$$\Pr(\bigcup_{j=1}^{K-1} E_j) \to 0 \text{ as } n \to \infty \tag{10}$$

Proof. Let $(X_1,Y_1),(X_2,Y_2) \sim p_{X,Y}$ be two pairs of random variables. We define

$$p_0 \triangleq \Pr(Y_1 \neq Y_2 | X_1 \perp\!\!\!\perp X_2) \tag{11}$$

$$p_1 \triangleq \Pr(Y_1 \neq Y_2 | X_1 = X_2) \tag{12}$$

Observe that Y_1 and Y_2 are noisy observations of independent database entries X_1, X_2 when $X_1 \perp\!\!\!\perp X_2$ and Y_1 and Y_2 are noisy replicas when $X_1 = X_2$. We can rewrite p_0 and p_1 as the following.

$$p_0 = \sum_{x_1\in\mathfrak{X}}\sum_{x_2\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} p_X(x_1)p_X(x_2)p_{Y|X}(y|x_1)\left[1-p_{Y|X}(y|x_2)\right] \tag{13}$$

$$= \sum_{x_1\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} p_X(x_1)p_{Y|X}(y|x_1)\sum_{x_2\in\mathfrak{X}} p_X(x_2)\left[1-p_{Y|X}(y|x_2)\right] \tag{14}$$

$$= \sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} p_X(x)p_{Y|X}(y|x)\left[1-p_Y(y)\right] \tag{15}$$

$$p_1 = \sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} p_X(x)p_{Y|X}(y|x)\left[1-p_{Y|X}(y|x)\right] \tag{16}$$

Thus, we have

$$p_0 - p_1 = \sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} p_{X,Y}(x,y)\left[p_{Y|X}(y|x) - p_Y(y)\right] \tag{17}$$

For every $y \in \mathfrak{X}$, let

$$\psi(y) \triangleq \sum_{x\in\mathfrak{X}} p_X(x)\left[p_{Y|X}(y|x) - p_Y(y)\right]^2 \tag{18}$$

$$= \sum_{x\in\mathfrak{X}} p_X(x)\left[p_{Y|X}(y|x) - \sum_{z\in\mathfrak{X}} p_{Y|X}(y|z)p_X(z)\right]^2 \tag{19}$$

$$\geq 0 \tag{20}$$

where (20) follows from the non-negativity of the square term in the summation. It must be noted that $\psi(y) = 0$ only if $p_{Y|X}(y|x) = p_Y(y) \forall x \in \mathfrak{X}$ with $p_X(x) > 0$.

Now, expanding the square term, we obtain

$$\psi(y) = \sum_{x\in\mathfrak{X}} p_X(x)p_{Y|X}(y|x)^2 - 2p_Y(y)\sum_{x\in\mathfrak{X}} p_X(x)p_{Y|X}(y|x) + \sum_{x\in\mathfrak{X}} p_X(x)p_Y(y)^2 \tag{21}$$

$$= \sum_{x\in\mathfrak{X}} p_X(x)p_{Y|X}(y|x)^2 - 2p_Y(y)^2 + p_Y(y)^2 \tag{22}$$

$$= \sum_{x\in\mathfrak{X}} p_X(x)p_{Y|X}(y|x)^2 - p_Y(y)^2 \tag{23}$$

Now, we rewrite $p_0 - p_1$ as

$$p_0 - p_1 = \sum_{y\in\mathfrak{X}}\sum_{x\in\mathfrak{X}} p_{X,Y}(x,y)\left[p_{Y|X}(y|x) - p_Y(y)\right] \tag{24}$$

$$= \sum_{y\in\mathfrak{X}}\left[\left(\sum_{x\in\mathfrak{X}} p_X(x)p_{Y|X}(y|x)^2\right) - p_Y(y)^2\right] \tag{25}$$

$$= \sum_{y\in\mathfrak{X}} \psi(y) \tag{26}$$

$$\geq 0 \tag{27}$$

with $p_0 - p_1 = 0$ only when $p_{Y|X}(y|x) = p_Y(y) \forall x,y \in \mathfrak{X}$. In other words, $p_0 > p_1$ as long as the two databases are not independent.

Choose any $\tau \in (p_1, p_0)$ bounded away from both p_0 and p_1. Let A_j denote the event that $C_j^{m_n}$ and $C_{j+1}^{m_n}$ are replicas and B_j denote the event that the algorithm detects $C_j^{m_n}$ and $C_{j+1}^{m_n}$ as replicas. From the union bound,

$$\Pr(\bigcup_{j=1}^{K-1} E_j) \leq \sum_{j=1}^{K-1} \Pr(A_j^c)\Pr(B_j|A_j^c) + \Pr(A_j)\Pr(B_j^c|A_j) \tag{28}$$

Note that conditioned on A_j^c, $d_H(C_j^{m_n}, C_{j+1}^{m_n}) \sim \text{Binom}(m_n, p_0)$ and conditioned on A_j, $d_H(C_j^{m_n}, C_{j+1}^{m_n}) \sim \text{Binom}(m_n, p_1)$. Then, from Chernoff bound [17, Theorem 1], we get

$$\Pr(B_j|A_j^c) \leq 2^{-m_n D(\tau\|p_0)} \tag{29}$$

$$\Pr(B_j^c|A_j) \leq 2^{-m_n D((1-\tau)\|1-p_1)} \tag{30}$$

where $D(.\|.)$ denotes the Kullback-Leibler divergence [18, Chapter 2.3] between two Bernoulli distributions with given parameters. Thus, we get

$$\Pr(\bigcup_{j=1}^{K-1} E_j) \leq (K-1)\left[2^{-m_n D(\tau\|p_0)} + (2^{-m_n D((1-\tau)\|1-p_1)}\right] \tag{31}$$

Observing that RHS of (28) has $2K-2 = O(n)$ terms decaying exponentially in m_n and $n \sim \log m_n$ concludes the proof. □

B. Deletion Detection Using Seeds

e propose to detect deletions using seeds. Let $(\mathbf{G}^{(1)}, \mathbf{G}^{(2)})$ be a batch of $B = \Theta(n^d)$ seeds. Our deletion detection algorithm works as follows: After finding the replicas as in Section III-A, we discard all-but-one of the noisy replicas from $\mathbf{G}^{(2)}$, to obtain $\tilde{\mathbf{G}}^{(2)}$ whose column size is denoted by $\tilde{K}$. At this step, we only have deletions.

We adopt an exhaustive search over all potential deletion patterns with $n-\tilde{K}$ deletions on $\mathbf{G}^{(1)}$. For each deletion pattern

I, we compute the total Hamming distance $d_H(\tilde{\mathbf{G}}^{(1)}(I), \tilde{\mathbf{G}}^{(2)})$ between $\tilde{\mathbf{G}}^{(1)}(I)$ and $\tilde{\mathbf{G}}^{(2)}$, where $\tilde{\mathbf{G}}^{(1)}(I)$ denotes the matrix obtained by discarding the columns whose indices lie in I from $\mathbf{G}^{(1)}$. More formally, we compute

$$d_H(\tilde{\mathbf{G}}^{(1)}(I), \tilde{\mathbf{G}}^{(2)}) = \sum_{i \in [m_n]} \sum_{j \in [n-\tilde{K}]} \mathbb{1}_{\left[\tilde{G}^{(1)}(I)_{i,j} \neq \tilde{G}^{(2)}_{i,j}\right]} \tag{32}$$

Then, the algorithm outputs the deletion pattern minimizing total Hamming distance between $\tilde{\mathbf{G}}^{(1)}(I)$ and $\tilde{\mathbf{G}}^{(2)}$, denoted by $\hat{I}_{\text{del}}$. In other words,

$$\hat{I}_{\text{del}} = \underset{I \subseteq [n], |I| = n-\tilde{K}}{\operatorname{argmin}} d_H(\tilde{\mathbf{G}}^{(1)}(I), \tilde{\mathbf{G}}^{(2)}) \tag{33}$$

Note that such a strategy depends on pairs of correlated entries in $\mathbf{G}^{(1)}$ and $\tilde{\mathbf{G}}^{(2)}$ having a higher probability of being equal than independent pairs. More formally, given a correlated pair $(X_1, Y_1) \sim p_{X,Y}$, and an independent pair $(X_2, Y_1) \sim p_X p_Y$ we need

$$\Pr(Y_1 = X_1) > \Pr(Y_1 = X_2) \tag{34}$$

which is not true in general.

For example, suppose $\mathfrak{X} = \{0,1\}$ with $p_X(0) = 1/2$ and $p_{Y|X}$ follows BSC(q), *i.e.* $p_{Y|X}(x|x) = 1-q$, $x = 0,1$. Note that when $q > 1/2$ (34) is not satisfied. However, we can flip the output bits, by applying the bijective remapping $\sigma = \left(\begin{smallmatrix} 1 & 2 \\ 2 & 1 \end{smallmatrix}\right)$ to Y in order to satisfy (34).

Thus, as long as such a bijective remapping $\sigma : \mathfrak{X} \to \mathfrak{X}$ satisfying (34) exists, we can use the aforementioned deletion detection algorithm. Now, suppose that such a mapping σ exists. We apply σ to the entries of $\tilde{\mathbf{G}}^{(2)}$ to construct $\tilde{\mathbf{G}}^{(2)}_{\sigma}$. Then, our deletion detection algorithm computes $d_H(\tilde{\mathbf{G}}^{(1)}(I), \tilde{\mathbf{G}}^{(2)}_{\sigma})$ for each potential deletion pattern I and outputs the pattern $\hat{I}_{\text{del}}(\sigma)$ minimizing it. In other words,

$$\hat{I}_{\text{del}}(\sigma) = \underset{I \subseteq [n], |I| = n-\tilde{K}}{\operatorname{argmin}} d_H(\tilde{\mathbf{G}}^{(1)}(I), \tilde{\mathbf{G}}^{(2)}_{\sigma}) \tag{35}$$

The following lemma states that such a bijective mapping σ exists and for a seed order $d \geq 1$, this algorithm can infer the deletion locations with high probability.

Lemma 2. ***(Seeded Deletion Detection)*** *For a repetition pattern S^n, let $I_{del} = \{j \in [n] | S_j = 0\}$. Then there exists a bijective mapping σ depending on $p_{X,Y}$ satisfying* (34) *and for seed order $d = 1$,*

$$\Pr\left(\hat{I}_{del}(\sigma) = I_{del}\right) \to 1 \text{ as } n \to \infty \tag{36}$$

Proof. We first prove the existence of such a bijective mapping σ, satisfying (34). For all σ, let

$$\begin{aligned} q_0(\sigma) &= \Pr(\sigma(Y_1) \neq X_2) \\ &\triangleq \sum_{x_1 \in \mathfrak{X}} \sum_{x_2 \in \mathfrak{X}} p_X(x_1) p_X(x_2) [1 - p_{Y|X}(\sigma^{-1}(x_2)|x_1)] \end{aligned} \tag{37}$$

$$\begin{aligned} q_1(\sigma) &\triangleq \Pr(\sigma(Y_1) \neq X_1) \\ &= \sum_{x \in \mathfrak{X}} p_X(x) [1 - p_{Y|X}(\sigma^{-1}(x)|x)] \end{aligned} \tag{38}$$

Here, our goal is to show that there exists at least one σ satisfying

$$q_0(\sigma) > q_1(\sigma) \tag{39}$$

We first prove

$$\sum_{\sigma} q_0(\sigma) - q_1(\sigma) = 0 \tag{40}$$

where the summation is over all permutations σ. For brevity, let

$$P_{i,j} \triangleq p_{Y|X}(j|i) \quad \forall i, j \in \mathfrak{X} \tag{41}$$

Note that from (41), we have

$$\sum_{j=1}^{|\mathfrak{X}|} P_{i,j} = 1 \quad \forall i \in \mathfrak{X} \tag{42}$$

$$\sum_{i=1}^{|\mathfrak{X}|} \sum_{j=1}^{|\mathfrak{X}|} P_{i,j} = |\mathfrak{X}| \tag{43}$$

Taking the sum over all σ, we obtain

$$\begin{aligned} \sum_{\sigma} q_0(\sigma) - q_1(\sigma) = \sum_{\sigma} \sum_{i=1}^{|\mathfrak{X}|} \sum_{j=1}^{|\mathfrak{X}|} p_X(i) p_X(j) P_{i,\sigma^{-1}(j)} \\ - \sum_{\sigma} \sum_{i=1}^{|\mathfrak{X}|} p_X(i) P_{i,\sigma^{-1}(i)} \end{aligned} \tag{44}$$

Combining (42)-(44), it can be shown that both terms on the RHS of (44) are equal to $(|\mathfrak{X}|-1)!$. Thus, we have proved (40).

Now, we only need to show that

$$\exists \sigma \quad q_0(\sigma) - q_1(\sigma) \neq 0 \tag{45}$$

Considering several one-cycle permutations over $\mathfrak{X}$, one can show that

$$q_0(\sigma) - q_1(\sigma) = 0 \ \ \forall \sigma \iff p_{Y|X}(y|x) = p_Y(y) \ \ \forall (x,y) \in \mathfrak{X}^2 \tag{46}$$

We have assumed the databases are not independent, *i.e.,* $p_{X,Y} \neq p_X p_Y$. Thus, there exists a bijective mapping σ satisfying (39).

Now choose such a mapping σ. Let $\hat{K} = \sum_{j=1}^{n} \mathbb{1}_{[S_j \neq 0]}$ and Λ_n be the seed size. Let $\varepsilon > 0$ and declare error if $\hat{K} \notin [(1-\delta-\varepsilon)n, (1-\delta+\varepsilon)n]$ whose probability is denoted by κ_n. Then, we use the union bound to obtain

$$\begin{aligned} \Pr\left(\hat{I}_{\text{del}}(\sigma) \neq I_{\text{del}}\right) \leq \kappa_n + \\ \sum_{I \subseteq [n], |I| = \hat{K}} \Pr(d_H(\tilde{\mathbf{G}}^{(1)}(I), \tilde{\mathbf{G}}^{(2)}_{\sigma}) \leq d_H(\tilde{\mathbf{G}}^{(1)}(I_{\text{del}}), \tilde{\mathbf{G}}^{(2)}_{\sigma})) \end{aligned} \tag{47}$$

where the difference of the total Hamming distances in (47) can be written as the difference of two Binomial random variables with a common number of trials depending on the size of the overlap between I and I_{del}.

Specifically, denote by $f(I, I_{\text{del}})$ the number of overlapping elements between $[n] \setminus I$ and $[n] \setminus I_{\text{del}}$. Here we count the overlaps as follows: We count $i_1 \in ([n] \setminus I) \bigcap [n] \setminus I_{\text{del}}$ as an overlapping element only if i_1 is in the same position in each

one of the ordered sets $i_1 \in ([n] \setminus I)$ and $[n] \setminus I_{\text{del}}$. For example, let $n = 3$, $I = \{1\}$, $I_{\text{del}} = \{3\}$. Then we have $[n] \setminus I = \{2,3\}$ and $[n] \setminus I_{\text{del}} = \{1,2\}$. Note that even though the element 2 is present in both sets, it is in different positions when the sets are ordered. In this case, we have $f(I, I_{\text{del}}) = 0$.

Now, observe that

$$d_H(\tilde{\mathbf{G}}^{(1)}(I), \tilde{\mathbf{G}}^{(2)}_{\sigma}) - d_H(\tilde{\mathbf{G}}^{(1)}(I_{\text{del}}), \tilde{\mathbf{G}}^{(2)}_{\sigma}) \tag{48}$$

can be written as the difference between two Binomial random variables with respective parameters $(\Lambda_n(\hat{K} - f(I, I_{\text{del}})), q_0(\sigma))$ and $(\Lambda_n(\hat{K} - f(I, I_{\text{del}})), q_1(\sigma))$. From Hoeffding's inequality [17], we obtain

$$\begin{aligned}
&\Pr(d_H(\tilde{\mathbf{G}}^{(1)}(I), \tilde{\mathbf{G}}^{(2)}_{\sigma}) \le d_H(\tilde{\mathbf{G}}^{(1)}(I_{\text{del}}), \tilde{\mathbf{G}}^{(2)}_{\sigma})) \\
&\quad = \Pr(d_H(\tilde{\mathbf{G}}^{(1)}(I), \tilde{\mathbf{G}}^{(2)}_{\sigma}) - d_H(\tilde{\mathbf{G}}^{(1)}(I_{\text{del}}), \tilde{\mathbf{G}}^{(2)}_{\sigma}) \le 0) && (49) \\
&\quad \le \exp\left(-\frac{1}{2}\Lambda_n(\hat{K} - f(I, I_{\text{del}}))(q_0(\sigma) - q_1(\sigma))^2\right) && (50) \\
&\quad = q^{\Lambda_n(\hat{K} - f(I, I_{\text{del}}))} && (51)
\end{aligned}$$

where

$$q \triangleq e^{-\frac{1}{2}(q_0(\sigma) - q_1(\sigma))^2} < 1 \tag{52}$$

Furthermore, the number of false deletion index sets I with a given $f(I, I_{\text{del}})$ can be wastefully upper bounded by $\binom{n}{\hat{K}}$. Thus, we can further bound the probability of error as

$$\begin{aligned}
\Pr\left(\hat{I}_{\text{del}}(\sigma) \ne I_{\text{del}}\right) &\le \kappa_n + \sum_{i=0}^{\hat{K}-1} \binom{n}{\hat{K}} q^{\Lambda_n(\hat{K}-i)} && (53) \\
&= \kappa_n + \binom{n}{\hat{K}} \sum_{i=0}^{\hat{K}-1} q^{\Lambda_n(\hat{K}-i)} && (54) \\
&= \kappa_n + \binom{n}{\hat{K}} \sum_{j=1}^{\hat{K}} q^{\Lambda_n j} && (55) \\
&= \kappa_n + \binom{n}{\hat{K}} \sum_{i=0}^{\hat{K}-1} q^{\Lambda_n(i+1)} && (56) \\
&= \kappa_n + \binom{n}{\hat{K}} \sum_{i=0}^{\hat{K}-1} q^{\Lambda_n} q^{\Lambda_n i} && (57) \\
&= \kappa_n + \binom{n}{\hat{K}} q^{\Lambda_n} \sum_{i=0}^{\hat{K}-1} q^{\Lambda_n i} && (58) \\
&\le \kappa_n + 2^{nH_b(\hat{K}/n)} q^{\Lambda_n} \frac{1 - q^{\Lambda_n \hat{K}}}{1 - q^{\Lambda_n}} && (59) \\
&\le \kappa_n + 2^{nH_b(\hat{K}/n)} q^{\Lambda_n} \frac{1}{1-q} && (60) \\
&= \kappa_n + \frac{1}{1-q} 2^{nH_b(\hat{K}/n) - \Lambda_n \log \frac{1}{q}} && (61)
\end{aligned}$$

where H_b denotes the binary entropy function. Observe that the RHS of (61) vanishes as $n \to \infty$ if

$$\Lambda_n \ge \frac{nH_b(\hat{K}/n)}{\log \frac{1}{q}} = \frac{2nH_b(\hat{K}/n)}{(q_0(\sigma) - q_1(\sigma))^2 \log e} \tag{62}$$

which can be satisfied with some $\Lambda_n = \Theta(n)$. Thus a seed order $d = 1$ is sufficient for successful deletion detection. □

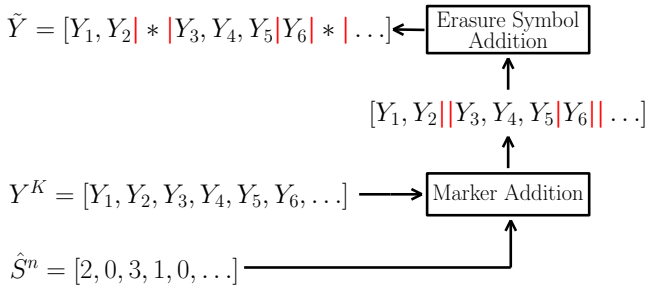

Fig. 3. An example of the construction of $\tilde{\mathbf{D}}^{(2)}$, as described in Step 3 of the proof of Theorem 1, illustrated over a pair of rows X^n of $\mathbf{D}^{(1)}$ and Y^K of $\mathbf{D}^{(2)}$. After these steps, in Step 4 we check the joint typicality of the rows X^n of $\mathbf{D}^{(1)}$ and $\tilde{Y}$ of $\tilde{\mathbf{D}}^{(2)}$.

In contrast with the linear seed size of Lemma 2, [11] requires that the number of seeds is logarithmic in the number of columns. This is because in [11] the performance criterion is the successful detection of an *arbitrarily-chosen* deleted column, whereas in this work, the criterion is the successful detection of *all* deleted columns.

C. Row Matching Scheme and Achievability

We are now ready to outline the proof of achievability of Theorem 1.

Proof of Achievability of Theorem 1. Let S^n be the underlying repetition pattern and $K \triangleq \sum_{i=1}^n S_i$ be the number of columns in $\mathbf{D}^{(2)}$. The matching scheme we propose follows these steps:

1) Perform replica detection as in Section III-A. The probability of error of this step is denoted by ρ_n.

2) Perform deletion detection using seeds as in Section III-B. The probability of error is denoted by μ_n. At this step, we have an estimate $\hat{S}^n$ of S^n.

3) Using $\hat{S}^n$, place markers between the noisy replica runs of different columns to obtain $\tilde{\mathbf{D}}^{(2)}$. If a run has length 0, *i.e.* deleted, introduce a column consisting of erasure symbol $* \notin \mathfrak{X}$. Note that provided that the detection algorithms in Steps 1 and 2 have performed correctly, there are exactly n such runs, where the j^{th} run in $\tilde{\mathbf{D}}^{(2)}$ corresponds to the noisy copies of the j^{th} column of $\Theta_n \circ \mathbf{D}^{(1)}$ if $S_j \ne 0$, and an erasure column otherwise.

4) Fix $\varepsilon > 0$. Match the l^{th} row Y_l^K of $\tilde{\mathbf{D}}^{(2)}$ with the i^{th} row X_i^n of $\mathbf{D}^{(1)}$, if X_i is the only row of $\mathbf{D}^{(1)}$ jointly ε-typical with Y_l^K according to $p_{X,Y^S,S}$, assigning $\hat{\Theta}_n(i) = l$, where

$$p_{X,Y^S|S}(x, y^s|s) = \begin{cases} p_X(x)\mathbb{1}_{[y^s = *]} & \text{if } s = 0 \\ p_X(x) \prod_{j=1}^{s} p_{Y|X}(y_j|x) & \text{if } s \ge 1 \end{cases} \tag{63}$$

with $y^s = y_1 \ldots y_s$. Otherwise, declare an error.

The column discarding and the marker addition as described in Steps 3-4, are illustrated in Figure 3.

The total probability of error of this scheme (as in (3)) can be bounded as follows

$$P_e \le 2^{nR} 2^{-n(I(X;Y^S,S) - 3\varepsilon)} + \varepsilon + \rho_n + \mu_n \tag{64}$$

Note that since m_n is exponential in n, $d \geq 1$, and from WLLN, using Lemma 1 we have $\rho_n \to 0$ and using Lemma 2 we have $\mu_n \to 0$ as $n \to \infty$. Thus $P_e \leq \varepsilon$ as $n \to \infty$ if $R < I(X;Y^S,S)$, concluding the proof. □

The matching scheme proposed above for noisy repeated database matching is different from the one proposed in [12] for the noiseless setting in several ways: First, in the noiseless setting, the seeds are not required and a single detection algorithm can identify deletions and replicas. Second, in Step 3 of the proof above, unlike [12], the noisy replicas are retained. This is because under noise, replicas offer additional information, similar to a repetition code. This implies an important distinction between database matching under synchronization errors and decoding in a repeat channel [19]: In database matching, the identical repetition pattern over a large number of rows allows us to detect deletions and replicas, which in turn improves the achievable database growth rate. On the other hand, in a repeat channel, detecting the repetition pattern is in general not possible and the replicas have a negative impact on the channel capacity.

D. Converse

We argue that the database growth rate achieved in Theorem 1 is in fact tight using a genie-aided proof through Fano's inequality where the repetition pattern S^n is known. We argue that since the rows are *i.i.d.* conditioned on the repetition pattern S^n, the seeds $(\mathbf{G}^{(1)},\mathbf{G}^{(2)})$ do not offer any additional information given S^n. Therefore, as the seeds become irrelevant in this genie-aided proof, we argue that the converse result holds for any seed order d.

Proof of Converse of Theorem 1. Let R be the database growth rate and P_e be the probability that the scheme is unsuccessful for a uniformly-selected row pair. More formally,

$$P_e \triangleq \Pr\left(\Theta_n(J) \neq \hat{\Theta}_n(J)\right), \quad J \sim \mathrm{Unif}([m_n]) \tag{65}$$

Furthermore, let S^n be the repetition pattern and $K = \sum_{j=1}^{n} S_j$. Since Θ_n is a uniform permutation, from Fano's inequality, we have

$$H(\Theta) \leq 1 + m_n P_e \log m_n + I(\Theta_n;\mathbf{D}^{(1)},\mathbf{D}^{(2)},\mathbf{G}^{(1)},\mathbf{G}^{(2)}) \tag{66}$$

From the independence of Θ_n, $\mathbf{D}^{(2)}$ and $(\mathbf{G}^{(1)},\mathbf{G}^{(2)})$, we get

$$\begin{aligned}
I(\Theta_n;\mathbf{D}^{(1)},\mathbf{D}^{(2)},\mathbf{G}^{(1)},\mathbf{G}^{(2)}) &= I(\Theta_n;\mathbf{D}^{(1)}|\mathbf{D}^{(2)},\mathbf{G}^{(1)},\mathbf{G}^{(2)}) && (67)\\
&\leq I(\Theta_n,\mathbf{D}^{(2)},\mathbf{G}^{(1)},\mathbf{G}^{(2)};\mathbf{D}^{(1)}) && (68)\\
&\leq I(\Theta_n,\mathbf{D}^{(2)},S^n;\mathbf{D}^{(1)}) && (69)\\
&= m_n I(Y^K,S^n;X^n) && (70)\\
&= m_n n I(X;Y^S,S) && (71)
\end{aligned}$$

where (69) follows from the fact that given S^n, $\mathbf{G}^{(1)},\mathbf{G}^{(2)}$ do not offer any additional information. Equation (70) follows from the fact that non-matching rows are *i.i.d.* conditioned on the repetition pattern S^n. Furthermore, (71) follows from the fact that the entries of $\mathbf{D}^{(1)}$ *i.i.d.*, and the noise on the entries are also *i.i.d.*

Finally, from Stirling's approximation and (71), we obtain

$$\begin{aligned}
R &= \lim_{n\to\infty} \frac{1}{m_n n} H(\Theta_n) && (72)\\
&\leq \lim_{n\to\infty} \left[\frac{1}{m_n n} + P_e R + I(X;Y^S,S)\right] && (73)\\
&\leq I(X;Y^S,S) && (74)
\end{aligned}$$

where (74) follows from the fact that $P_e \to 0$ as $n \to \infty$. □

IV. Conclusion

In this work, we have studied the database matching problem under random noisy column repetitions. We have showed that the running Hamming distances between the consecutive columns of the labeled noisy repeated database can be used to detect replicas. In addition, given seeds whose size grows logarithmic with the number of rows, an exhaustive search over the deletion patterns can be used to infer the locations of the deletions. Using the proposed detection algorithms, and a joint typicality based rowwise matching scheme, we have derived an achievable database growth rate, which we prove is tight. Therefore, we have completely characterized the database matching capacity under noisy column repetitions.

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] 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.

[9] 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.

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

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

[12] S. Bakirtas and E. Erkip, "Matching of markov databases under random column repetitions," *arXiv*, vol. abs/2202.01730, 2022. [Online]. Available: http://arxiv.org/abs/2202.01730

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

[14] Y. Li and G. Han, "Input-constrained erasure channels: Mutual information and capacity," in *Proc. of IEEE International Symposium on Information Theory (ISIT)*, 2014, pp. 3072–3076.

[15] 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.
[16] 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.
[17] W. Hoeffding, "Probability inequalities for sums of bounded random variables," in *The collected works of Wassily Hoeffding*. Springer, 1994, pp. 409–426.
[18] T. M. Cover, *Elements of Information Theory*. John Wiley & Sons, 2006.
[19] 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.