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.**

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 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 the anonymized and obfuscated databases are prone to privacy attacks.

Recently, matching correlated pairs of databases have been investigated from an information-theoretic point of view in [8]–[10]. In [9], 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 probability of error for all users. In [8], 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

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

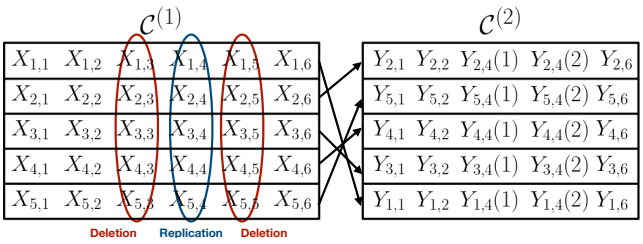

Fig. 1. An illustrative example of database matching under noisy column repetitions. The columns circled in red are deleted whereas the fourth column, which is circled in blue, is repeated twice, *i.e.,* replicated. For each i, $Y_{i,4}(1)$ and $Y_{i,4}(2)$ are noisy replicas of $X_{i,4}$. Our goal is to estimate the row permutation $\Theta = \left(\begin{smallmatrix}1&2&3&4&5\\5&1&4&3&2\end{smallmatrix}\right)$, by matching the rows of $\mathscr{C}^{(1)}$ and $\mathscr{C}^{(2)}$. Here the i^{th} row of $\mathscr{C}^{(1)}$ corresponds to the $\Theta(i)^{\text{th}}$ row of $\mathscr{C}^{(2)}$.

sufficient and necessary conditions on the *database growth rate* for reliable matching, in the presence of noise on the database entries.

In [10], 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 of different number of attributes (columns), unlike [8] where the row and the column sizes of the databases are the same. 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 a companion paper [11], we investigate database matching under noiseless column repetition, a non-trivial extension of [10] to *column repetitions*. Under this generalized model, we devise a *histogram-based* repetition detection algorithm and derive an improved achievable rate. We then prove a converse showing the tightness of this achievable rate, thereby characterizing the exact matching capacity of database matching under column repetitions.

In this paper, our goal is to investigate the necessary and the sufficient conditions for the successful matching of rows under *noisy* column repetitions. We assume a generalized database model where synchronization errors, in the form of column repetitions, are followed by obfuscation, in the form

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

of independent noise on the database entries. The presence of noise prevents us from using the histogram-based repetition detection algorithm of [11] and unlike [11] requires *seed* users whose identities are known in both databases [10], [12], [13]. 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 logarithmic in the number of rows m of the database, deletion locations can be extracted from the seeds. Then, we propose a joint typicality-based matching scheme to derive sufficient for the 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 the proof of 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. 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 adapted from [8], [10], [11] 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 $\mathfrak{X}=\{1,\ldots,|\mathfrak{X}|\}$.

In order to characterize the relationship between m and n, we use the *database growth rate* introduced in [8].

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

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

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 $\{0,\ldots,s_{\max}\}$.

In this paper we assume that $\mathbf{S}^n$ and $\mathscr{C}^{(1)}$ are independent. According to Definiton 3, 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 $\mathscr{C}^{(1)}$ is said to be *deleted* and when $S_i>1$, i^{th} column of $\mathscr{C}^{(1)}$ is said to be *replicated*.

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 the *noise* $p_{Y|X}$ 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:

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

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 $\mathscr{C}^{(2)}$, as described in Definiton 4, 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 row $\Theta^{-1}(i)$. 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*.

Note that (3) states $\mathscr{C}^{(2)}_{i,j}$ 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. We stress that $p_{Y|X}$ is a general DMC model, capturing any distortion and noise on the database entries, though we only refer to this as "noise" in this paper.

In the noisy setting, inferring the column repetition pattern is a harder task, compared to the noiseless setting investigated in [11]. Therefore, we assume the availability of *seeds*, as done in database matching [10] and graph matching [12], [13] literatures.

Definition 5. (Seeds) For the unlabeled and labeled databases in Definitions 1 and 4, a *seed* is a pair of correctly-matched rows $(\mathbf{X},\mathbf{Y})$. A *batch of B seeds* $(\mathscr{D}^{(1)},\mathscr{D}^{(2)})$ is a pair of $B\times n$ and $B\times\sum_{j=1}^{n}S_j$ submatrices. Here, we assume that the *seed size* $B=\Theta(n^d)$ where d is called the *seed order*.

Definition 6. (Successful Matching Scheme) Given a seed order d, 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

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

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

Definition 7. (Achievable Database Growth Rate) Given a database probability distribution p_X, a repetition probability

distribution p_S, a noise distribution $p_{Y|X}$ and a seed order d, 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, a noise distribution $p_{Y|X}$ and a seed order d, the *matching capacity* $C(d)$ is the supremum of the set of all achievable rates.

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. Note that this generalizes [11] where we have no noise $(p_{Y|X}(y|x) = \mathbb{1}_{y=x}\,\forall x,y)$ and $B = 0$.

III. Main Result and Achievability

In this section, we present our main result on the matching capacity, stated in the following theorem and prove its achievability by proposing deletion and replica detection and database matching algorithms. We prove the converse in Section IV.

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

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

where $\mathbf{Y}(X,S) = Y_1,\ldots,Y_S$ such that

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

Theorem 1 states that although the repetition pattern is not known a-priori, given a seed order $d \geq 1$, we can achieve the above database growth rate which assumes a-priori knowledge of the repetition pattern. Since a higher seed order d would facilitate the matching, we will focus on seed order $d = 1$, which we show is sufficient to achieve the matching capacity.

As we discuss in [11], in the noiseless setting $\mathbf{Y}(X,S) = X \otimes \mathbf{1}_{1\times S}$, and in turn, the replicas do not offer any additional information. Hence we have the following corollary.

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

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

we have

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

for any seed order $d \geq 1$, where $\delta \triangleq p_S(0)$ is the deletion probability.

Corollary 1 states that in the noiseless setting, the main result of this paper coincides with that of [11], for any seed order $d \geq 1$. However, note that in the noiseless setting, the histogram-based detection algorithm of [11] does not require any seeds. Therefore, the result of Corollary 1 is in fact valid for any seed order d, including $d = -\infty$ indicating the case in which there are no seeds.

Corollary 2. ***(No Synchronization Errors)*** *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 $d \geq 1$.

Corollary 2 agrees with the main result of [8], restricted to any seed order $d \geq 1$. In fact, when there are no synchronization errors, seeds do not offer any information and hence, the result of Corollary 2 is in fact valid for any seed order d.

The rest of this section is on the achievability proof of Theorem 1. 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 its asymptotic performance. In Section III-B, we introduce a seeded deletion detection algorithm and derive a seed size sufficient for an asymptotic performance guarantee. Finally, in Section III-C, we combine these two algorithms and prove the achievability of Theorem 1 by generalizing the rowwise matching scheme proposed in [11] to the noisy scenario.

Throughout this section, we assume that the two databases are not independent. More formally, we assume

$$\exists (x,y) \in \mathfrak{X}^2 \quad p_{X,Y}(x,y) \neq p_X(x)p_Y(y) \tag{10}$$

since in order to perform matching, we need the labeled database to carry information on the unlabeled database. Note that when the two databases are independent, Theorem 1 states that the matching capacity becomes zero, hence Theorem 1 trivially holds.

A. Noisy Replica Detection

Given $\mathscr{C}^{(1)}$ and $\mathscr{C}^{(2)}$, we detect the replicas by extracting permutation-invariant features of the columns, similar to [11]. In [11], we choose the histogram of each column as its permutation-invariant feature and prove that these histograms are asymptotically unique. Then we look for exact matching between the histograms of $\mathscr{C}^{(1)}$ and $\mathscr{C}^{(2)}$ to infer the repetition process. 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 the arbitrary pairs of histograms are likely to be jointly typical, even though the columns are independent. Therefore, we propose a replica detection algorithm which adopts the Hamming distance between consecutive columns of $\mathscr{C}^{(2)}$ as the permutation-invariant feature.

Let K denote the number of columns of $\mathscr{C}^{(2)}$, $\mathbf{R}_j$ denote the i^{th} column of $\mathscr{C}^{(2)}$, $j = 1,\ldots,K$. Our replica detection algorithm works as follows: We first compute the Hamming distances $d_H(\mathbf{R}_i,\mathbf{R}_{i+1})$ between $\mathbf{R}_i$ and $\mathbf{R}_{i+1}$, for $i \in [K-1]$. For some threshold τ, the algorithm concludes that $\mathbf{R}_i$ and $\mathbf{R}_{i+1}$ are replicas if $d_H(\mathbf{R}_i,\mathbf{R}_{i+1}) < m\tau$, and independent

otherwise. We show that this algorithm can infer the replicas with high probability, as stated in the following lemma.

Lemma 1. ***(Noisy Replica Detection)*** *Consider the database matching problem of Theorem 1 where the two databases are not independent. Let E_i denote the event that the Hamming distance based algorithm described above fails to infer the correct relationship between $\mathbf{R}_i$ and $\mathbf{R}_{i+1}$, $i = 1,\dots,K-1$. Then*

$$\Pr\left(\bigcup_{i=1}^{K-1} E_i\right) \to 0 \text{ as } n \to \infty \tag{11}$$

where the average Hamming distance threshold τ is chosen based on $p_{X,Y}$.

Proof. See Appendix A. □

B. Seeded Deletion Detection

Let $(\mathscr{D}^{(1)}, \mathscr{D}^{(2)})$ be a batch of $B = \Theta(n^d)$ seeds, with 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 using the 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, we only have deletions.

After obtaining $\tilde{\mathscr{D}}^{(2)}$, we adopt a Hamming distance based strategy. However, note that such a strategy depends on pairs of correlated entries in $\mathscr{D}^{(1)}$ and $\tilde{\mathscr{D}}^{(2)}$ being more likely to be 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

$$\Pr(Y_1 = X_1) > \Pr(Y_1 = X_2) \tag{12}$$

which is not true in general.

For example, suppose $\mathfrak{X} = \{0,1\}$ with $p_X(0) = 1/2$ and BSC(q) noise distribution, *i.e.* $p_{Y|X}(x|x) = 1-q$, $x = 0,1$. Note that, when $q < 1/2$, (12) is automatically satisfied and when $q > 1/2$ (12) is not satisfied. However, we can flip the bits in $\tilde{\mathscr{D}}^{(2)}$, by applying the bijective remapping $\Phi = \left(\begin{smallmatrix} 1 & 2 \\ 2 & 1 \end{smallmatrix}\right)$ to satisfy (12).

Thus, as long as such a bijective remapping $\Phi : \mathfrak{X} \to \mathfrak{X}$ satisfying (12) exists, we can use a Hamming distance based deletion detection algorithm. Now, suppose that such a mapping Φ exists. We perform Φ to construct $\tilde{\mathscr{D}}^{(2)}_{\Phi}$. Then, we do an exhaustive search over all potential deletion patterns and for each deletion pattern I we compute the total Hamming distance between $\mathscr{D}^{(1)}([n] \setminus I)$ and $\tilde{\mathscr{D}}^{(2)}_{\Phi}$, where $\mathscr{D}^{(1)}([n] \setminus I)$ denotes the matrix obtained by discarding the columns whose indices lie in I. We output the deletion pattern minimizing total Hamming distance between $\mathscr{D}^{(1)}([n] \setminus I)$ and $\tilde{\mathscr{D}}^{(2)}_{\Phi}$, denoted by $\hat{I}_{\text{del}}$. In other words,

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

The following lemma states that given that (10) is satisfied, such a bijective mapping Φ exists and for a seed order $d \geq 1$, this algorithm can infer the deletion locations with high probability.

Lemma 2. ***(Seeded Deletion Detection)*** *Consider a joint distribution $p_{X,Y}$ satisfying* (10)*,* i.e., *the resulting random variables X and Y are not independent. For a repetition pattern $\mathbf{S}^n$, let $I_{del} = \{j \in [n] | S_j = 0\}$. Then there exists a bijective mapping Φ depending on $p_{X,y=Y}$ satisfying* (12) *and for seed order $d = 1$,*

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

Proof. See Appendix B. □

Note that since the seed order $d = 1$ is sufficient to infer the deletion locations and the additional seeds would facilitate deletion detection, the deletion detection algorithm described above works for any seed order $d \geq 1$.

The reason for the difference between the linear seed size in Lemma 2 and the logarithmic seed size in [10] is that although the detection algorithm of [10] can be adapted to the noisy setting, in [10] the performance criterion is successful detection of an arbitrarily-chosen deleted column, whereas in this work, the criterion is the successful detection of all deleted columns.

C. Matching Scheme

We are now ready to prove the achievability of Theorem 1.

Proof of Achievability of Theorem 1. We 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 mismatch can be made arbitrarily small asymptotically. 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 seeded deletion detection as in Section III-B. The probability of error is denoted by μ_n. At this step, we have an estimate $\hat{\mathbf{S}}^n$ of $\mathbf{S}^n$.
3) Using the estimate $\hat{\mathbf{S}}^n$ obtained in Step 2, discard the deleted columns from $\mathscr{C}^{(1)}$, to obtain $\tilde{\mathscr{C}}^{(1)}$, whose column size denoted by $\tilde{K} = n - \sum_{i=1}^{n} \mathbb{1}_{[S_i=0]}$.
4) Place markers between the noisy replica runs of different columns to obtain $\tilde{\mathscr{C}}^{(2)}$. Note that provided that the detection algorithms in Steps 1 and 2 have 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 $\Theta \circ \tilde{\mathscr{C}}^{(1)}$.
5) Fix $\varepsilon > 0$. Match the l^{th} row $\mathbf{Y}_l$ of $\tilde{\mathscr{C}}^{(2)}$ with the i^{th} row $\mathbf{X}_i$ of $\tilde{\mathscr{C}}^{(1)}$, if $\mathbf{X}_i$ is the only row of $\tilde{\mathscr{C}}^{(1)}$ jointly ε-typical with $\mathbf{Y}_l$ according to $p_{X,\mathbf{Y}(X,S),S}$, assigning $\hat{\Theta}(i) = l$, where

$$p_{X,\mathbf{Y}(X,S)|S}(x, \mathbf{y}(x,s)|S = s) = p_X(x) \prod_{j=1}^{s} p_{Y|X}(y_j|x) \tag{15}$$

with $\mathbf{y}(x,s) = y_1 \dots y_s$. Otherwise, declare an error.

Let $A_{\varepsilon}^{(n)}$ denote the set of jointly ε-typical sequence pairs according to $p_{X,\mathbf{Y}(X,S),S}$. For the matching rows $\mathbf{X}_1$,$\mathbf{Y}$ of $\tilde{\mathscr{C}}^{(1)}$

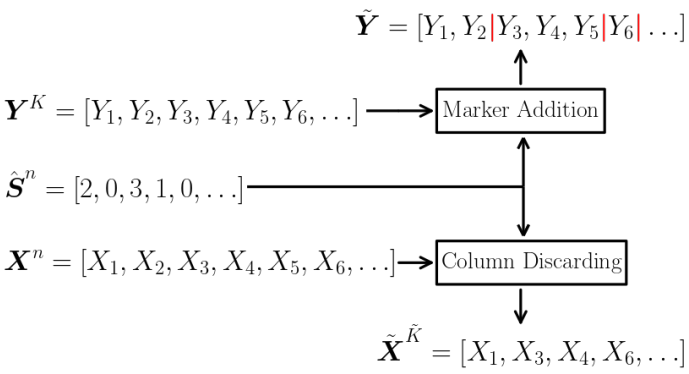

Fig. 3. An example of the construction of $\tilde{\mathscr{C}}^{(1)}$ and $\tilde{\mathscr{C}}^{(2)}$, as described in Steps 3 and 4, illustrated over a pair of rows $\mathbf{X}^n$ of $\mathscr{C}^{(1)}$ and $\mathbf{Y}^K$ of $\mathscr{C}^{(2)}$. After these steps, in Step 5 we check the joint typicality of the row $\tilde{X}^{\tilde{K}}$ of $\tilde{\mathscr{C}}^{(1)}$ and $\tilde{Y}$ of $\tilde{\mathscr{C}}^{(2)}$.

and $\tilde{\mathscr{C}}^{(2)}$, define the pairwise collision probability between $\mathbf{X}_1$ and $\mathbf{X}_i$ as

$$P_{\text{col,i}} \triangleq \Pr((\mathbf{X}_i, \mathbf{Y}) \in A_\varepsilon^{(n)} | (\mathbf{X}_1, \mathbf{Y}) \in A_\varepsilon^{(n)}) \, \forall i \in [m] \setminus \{1\}) \quad (16)$$

Note that since the rows are *i.i.d.*, we have

$$P_{\text{col,i}} = \Pr((\mathbf{X}_i, \mathbf{Y}) \in A_\varepsilon^{(n)}) \quad (17)$$

$$\leq 2^{-n(I(X;\mathbf{Y}(X,S),S)-3\varepsilon)} \quad (18)$$

Using the union bound, the total probability P_e of error of this scheme can be bounded as

$$P_e \leq \rho_n + \mu_n + \Pr((\mathbf{X}_1, \mathbf{Y}) \notin A_\varepsilon^{(n)}) + (1 - \Pr((\mathbf{X}_1, \mathbf{Y}) \notin A_\varepsilon^{(n)})) \sum_{i=2}^{2^{nR}} P_{col,i} \quad (19)$$

$$\leq \rho_n + \mu_n + \varepsilon + 2^{nR} P_{\text{col},2} \quad (20)$$

where we used the fact that the rows are *i.i.d.* and thus $P_{col,i} = P_{col,2} \ \forall i \in [m] \setminus \{1\}$.

Combining (18) and (20), we get

$$P_e \leq 2^{nR} 2^{-n(I(X;\mathbf{Y}(X,S),S)-3\varepsilon)} + \varepsilon + \rho_n + \mu_n \quad (21)$$

Note that since m is exponential in n, $d \geq 1$, and from WLLN, we have $\rho_n \to 0$, $\mu_n \to 0$ and $\kappa_n \to 0$ as $n \to \infty$. Thus,

$$P_e \leq \varepsilon \text{ as } n \to \infty \quad (22)$$

if $R < I(X;\mathbf{Y}(X,S),S) - 3\varepsilon$. Thus, we can argue that any positive rate R satisfying

$$R < I(X;\mathbf{Y}(X,S),S) \quad (23)$$

can be achieved by taking ε small enough. □

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

The matching scheme proposed above for noisy repeated database matching is different from the one proposed in [11] 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 4 of the proof above, unlike [11], the noisy replicas are retained. This is because under noise, replicas offer additional information, similar to a repetition code. However, there is an important distinction between database matching and decoding in a repeat channel: 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 not possible and the replicas have a negative impact on the channel capacity.

IV. Converse

In this section, we show that the database growth rate achieved in Theorem 1 is in fact tight.

Proof of Converse of Theorem 1. We prove the converse using the modified Fano's inequality presented in [8]. Let P_e be the probability that a given scheme does not identify a pair of matching rows. Since Θ is a uniform permutation, from Fano's inequaliy, we have

$$\frac{1}{mn} H(\Theta) \leq \frac{1}{mn} + \frac{1}{n} P_e \log m + \frac{1}{mn} I(\Theta; \mathscr{C}^{(1)}, \mathscr{C}^{(2)}, \mathscr{D}^{(1)}, \mathscr{D}^{(2)}) \quad (24)$$

$$I(\Theta; \mathscr{C}^{(1)}, \mathscr{C}^{(2)}, \mathscr{D}^{(1)}, \mathscr{D}^{(2)}) = I(\Theta; \mathscr{C}^{(1)} | \mathscr{C}^{(2)}, \mathscr{D}^{(1)}, \mathscr{D}^{(2)}) \quad (25)$$

Furthermore, we have

$$I(\Theta; \mathscr{C}^{(1)}, \mathscr{C}^{(2)}, \mathscr{D}^{(1)}, \mathscr{D}^{(2)}) = I(\Theta; \mathscr{C}^{(1)}, \mathscr{C}^{(2)} | \mathscr{D}^{(1)}, \mathscr{D}^{(2)}) \quad (26)$$

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

$$I(\Theta; \mathscr{C}^{(1)}, \mathscr{C}^{(2)} | \mathscr{D}^{(1)}, \mathscr{D}^{(2)}) = I(\Theta; \mathscr{C}^{(1)} | \mathscr{C}^{(2)}, \mathscr{D}^{(1)}, \mathscr{D}^{(2)}) \quad (27)$$

Now we can further upper bound the RHS of (27) as

$$I(\Theta; \mathscr{C}^{(1)} | \mathscr{C}^{(2)}, \mathscr{D}^{(1)}, \mathscr{D}^{(2)}) \leq I(\Theta, \mathscr{C}^{(2)}, \mathscr{D}^{(1)}, \mathscr{D}^{(2)}; \mathscr{C}^{(1)}) \quad (28)$$

$$\leq I(\Theta, \mathscr{C}^{(2)}, \mathbf{S}^n; \mathscr{C}^{(1)}) \quad (29)$$

$$= mI(\mathbf{X}^n; \mathbf{Y}, \mathbf{S}^n) \quad (30)$$

$$= mnI(X; \mathbf{Y}(X,S), S) \quad (31)$$

where (29) follows from the fact that given $\mathbf{S}^n$, $\mathscr{D}^{(1)}, \mathscr{D}^{(2)}$ do not offer any additional information. Equation (30) follows from the fact that non-matching rows are *i.i.d.* conditioned on the repetition pattern $\mathbf{S}^n$. Furthermore, (31) follows from the fact that the entries of $\mathscr{C}^{(1)}$ *i.i.d.*, and the noise on the entries are *i.i.d.*

We observe that, since we assumed the availability of $\mathbf{S}^n$ in (29), this bound holds for any seed size B.

So far, we have

$$\frac{1}{m} H(\Theta) \leq \frac{1}{m} + P_e \log m + nI(X; \mathbf{Y}(X,S), S) \quad (32)$$

Using Stirling's approximation

$$H(\Theta) = \log m! = m \log m - m \log e + O(\log m) \quad (33)$$

$$\lim_{n \to \infty} \frac{1}{mn} H(\Theta) = R \quad (34)$$

Using (34), we get

$$\frac{1}{mn}H(\Theta) \leq \frac{1}{mn} + P_e \frac{1}{n}\log m + I(X;\mathbf{Y}(X,S),S) \tag{35}$$

$$\lim_{n\to\infty}\frac{1}{mn}H(\Theta) \leq \lim_{n\to\infty}\left[\frac{1}{mn} + P_e R + I(X;\mathbf{Y}(X,S),S)\right] \tag{36}$$

$$R \leq I(X;\mathbf{Y}(X,S),S) \tag{37}$$

where (37) 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. 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 of size logarithmic in 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. Our ongoing work includes investigating the matching capacity under noisy repetitions 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. Bakirtas and E. Erkip, "Database matching under column repetitions," Available at: https://serhatbakirtas.github.io/files/noiselessdbmatching.pdf.

[12] 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.

[13] 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.

[14] W. Hoeffding, "Probability inequalities for sums of bounded random variables," in *The collected works of Wassily Hoeffding*. Springer, 1994, pp. 409–426.

Appendix

A. Proof of Lemma 1

For brevity, let $\rho_n = \Pr\left(\bigcup_{i=1}^{K-1} E_i\right)$ and $(X_1,Y_1),(X_2,Y_2) \sim p_{X,Y}$ be two pairs of random variables.

Now, let H_0 denote the event that X_1 and X_2 are independent and H_1 denote the event that $X_1 = X_2$. Let

$$p_0 \triangleq \Pr(Y_1 \neq Y_2 | H_0) \tag{38}$$

$$p_1 \triangleq \Pr(Y_1 \neq Y_2 | H_1) \tag{39}$$

We can rewrite p_0 and p_1 as the following.

$$p_0 = \sum_{x_1\in\mathfrak{X}}\sum_{x_2\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} p_X(x_1)p_X(x_2)p_{Y|X}(y|x_1)\left[1-p_{Y|X}(y|x_2)\right] \tag{40}$$

$$= \sum_{x_1\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} p_X(x_1)p_{Y|X}(y|x_1)\sum_{x_2\in\mathfrak{X}} p_X(x_2)\left[1-p_{Y|X}(y|x_2)\right] \tag{41}$$

$$= \sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} p_X(x)p_{Y|X}(y|x)\left[1-p_Y(y)\right] \tag{42}$$

$$p_1 = \sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} p_X(x)p_{Y|X}(y|x)\left[1-p_{Y|X}(y|x)\right] \tag{43}$$

Thus, we have

$$p_0 - p_1 = \sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} p_{X,Y}(x,y)\left[p_{Y|X}(y|x) - p_Y(y)\right] \tag{44}$$

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

$$\psi(y) \triangleq \sum_{x\in\mathfrak{X}} p_X(x)\left[p_{Y|X}(y|x) - p_Y(y)\right]^2 \tag{45}$$

$$= \sum_{x\in\mathfrak{X}} p_X(x)\left[p_{Y|X}(y|x) - \sum_{z\in\mathfrak{X}} p_{Y|X}(y|z)p_X(z)\right]^2 \tag{46}$$

$$\geq 0 \tag{47}$$

where (47) 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

$$\psi(y) = \sum_{x\in\mathfrak{X}} p_X(x)p_{Y|X}(y|x)^2 - 2p_Y(y)\sum_{x\in\mathfrak{X}} p_X(x)p_{Y|X}(y|x) + \sum_{x\in\mathfrak{X}} p_X(x)p_Y(y)^2 \tag{48}$$

$$= \sum_{x\in\mathfrak{X}} p_X(x)p_{Y|X}(y|x)^2 - 2p_Y(y)^2 + p_Y(y)^2 \tag{49}$$

$$= \sum_{x\in\mathfrak{X}} p_X(x)p_{Y|X}(y|x)^2 - p_Y(y)^2 \tag{50}$$

Now, we rewrite $p_0 - p_1$ as

$$p_0 - p_1 = \sum_{y\in\mathfrak{X}}\sum_{x\in\mathfrak{X}} p_{X,Y}(x,y)\left[p_{Y|X}(y|x) - p_Y(y)\right] \tag{51}$$

$$= \sum_{y\in\mathfrak{X}}\left[\left(\sum_{x\in\mathfrak{X}} p_X(x)p_{Y|X}(y|x)^2\right) - p_Y(y)^2\right] \tag{52}$$

$$= \sum_{y\in\mathfrak{X}} \psi(y) \tag{53}$$

$$\geq 0 \tag{54}$$

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, *i.e.,* (10) is satisfied.

Now if $\mathbf{R}_i$ and $\mathbf{R}_{i+1}$ are independent columns, the Hamming distance between $\mathbf{R}_i$ and $\mathbf{R}_{i+1}$ $d_H(\mathbf{R}_i,\mathbf{R}_{i+1})$ follows a $Binom(m,p_0)$ distribution whereas if $\mathbf{R}_i$ and $\mathbf{R}_{i+1}$ are noisy copies of each other, $d_H(\mathbf{R}_i,\mathbf{R}_{i+1})$ follows a $Binom(m,p_1)$ distribution.

After establishing $p_0 > p_1$, we choose the threshold as $\bar{\tau} = m\tau$, where $\tau \in (p_0,p_1)$ is bounded away from both p_0 and p_1. Let A_i denote the event that $\mathbf{R}_i$ and $\mathbf{R}_{i+1}$ are replicas and B_i denote the event that the algorithm detects $\mathbf{R}_i$ and $\mathbf{R}_{i+1}$ as replicas. Using the union bound, we can bound the error probability of this algorithm as

$$\rho_n \leq \sum_{i=1}^{K-1} \Pr(E_i) \tag{55}$$

$$\rho_n \leq \sum_{i=1}^{K-1} \Pr(A_i^c)\Pr(B_i|A_i^c) + \Pr(A_i)\Pr(B_i^c|A_i) \tag{56}$$

$$= \sum_{i=1}^{K-1} \Pr(A_i^c)\Pr(d_H(\mathbf{R}_i,\mathbf{R}_{i+1}) < m\tau|A_i^c) + \Pr(A_i)\Pr(d_H(\mathbf{R}_i,\mathbf{R}_{i+1}) \geq m\tau|A_i) \tag{57}$$

We, then can use the Chernoff bound to get

$$P(d_H(\mathbf{R}_i,\mathbf{R}_{i+1}) < m\tau|A_i^c) \leq e^{-mD(\tau\|p_0)} \tag{58}$$

$$P(d_H(\mathbf{R}_i,\mathbf{R}_{i+1}) \geq m\tau|A_i) \leq e^{-mD((1-\tau)\|1-p_1)} \tag{59}$$

From (57)-(59), we have

$$\rho_n \leq \sum_{i=1}^{K-1} \Pr(A_i^c)e^{-mD(\tau\|p_0)} + \Pr(A_i)e^{-mD((1-\tau)\|1-p_1)} \tag{60}$$

Thus as long as we choose τ bounded away from p_0 and p_1, we get $2K-2$ additive error terms, each decaying exponentially with m. Since m is exponential in n and $K \leq ns_{\max}$, $\rho_n \to 0$ as $n \to \infty$.

Therefore as long as (10) holds, we can choose an average Hamming distance threshold τ bounded away from p_0 and p_1, and detect any replicated column with an error probability decaying doubly-exponentially in the column size n. □

B. Proof of Lemma 2

We first prove the existence of such a bijective mapping Φ, given (12). For all Φ, let

$$q_0(\Phi) = \Pr(\Phi(Y_1) \neq X_2) \triangleq \sum_{x_1\in\mathfrak{X}}\sum_{x_2\in\mathfrak{X}} p_X(x_1)p_X(x_2)[1 - p_{Y|X}(\Phi^{-1}(x_2)|x_1)] \tag{61}$$

$$q_1(\Phi) \triangleq \Pr(\Phi(Y_1) \neq X_1) = \sum_{x\in\mathfrak{X}} p_X(x)[1 - p_{Y|X}(\Phi^{-1}(x)|x)] \tag{62}$$

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

$$q_0(\Phi) > q_1(\Phi) \tag{63}$$

We first prove

$$\sum_{\Phi} q_0(\Phi) - q_1(\Phi) = 0 \tag{64}$$

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

$$P_{i,j} \triangleq p_{Y|X}(j|i) \quad \forall i,j \in \mathfrak{X} \tag{65}$$

Note that from (65), we have

$$\sum_{j=1}^{|\mathfrak{X}|} P_{i,j} = 1 \quad \forall i \in \mathfrak{X} \tag{66}$$

$$\sum_{i=1}^{|\mathfrak{X}|}\sum_{j=1}^{|\mathfrak{X}|} P_{i,j} = |\mathfrak{X}| \tag{67}$$

Taking the sum over all Φ, we obtain

$$\sum_{\Phi} q_0(\Phi) - p_1(\Phi) = \sum_{\Phi}\sum_{i=1}^{|\mathfrak{X}|}\sum_{j=1}^{|\mathfrak{X}|} p_X(i)p_X(j)P_{i,\Phi^{-1}(j)} - \sum_{\Phi}\sum_{i=1}^{|\mathfrak{X}|} p_X(i)P_{i,\Phi^{-1}(i)} \tag{68}$$

Combining (66)-(68), it can be shown that both terms on the RHS of (68) are equal to $(|\mathfrak{X}|-1)!$. Thus, we have proved (64).

Now, we only need to show that

$$\exists\Phi \quad q_0(\Phi) - q_1(\Phi) \neq 0 \tag{69}$$

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

$$q_0(\Phi) - q_1(\Phi) = 0 \quad \forall\Phi \text{ iff } p_{Y|X}(y|x) = p_Y(y) \quad \forall(x,y) \in \mathfrak{X}^2 \tag{70}$$

We have assumed $p_{X,Y}$ satisfies (10). Thus, there exists a bijective mapping Φ satisfying (63).

Now, we perform replica detection via the algorithm proposed in Lemma 1 to detect all the replicas. Denote by ρ_n, the probability that the replica detection is performed incorrectly. By Lemma 1, $\rho_n \to 0$ as $n \to \infty$.

After detecting the replicas from $\mathscr{D}^{(2)}$, we discard all-but-one of the noisy replicas, to obtain $\tilde{\mathscr{D}}^{(2)}$. Then we perform the remapping Φ to obtain $\tilde{\mathscr{D}}^{(2)}_{\Phi}$.

Now, let $\hat{K}$ be the random variable corresponding to the number of retained columns in $\tilde{\mathscr{D}}_{\Phi}^{(2)}$. Note that, provided that the replica detection is done correctly, $\hat{K} \sim \text{Binom}(n, 1-\delta)$, and by WLLN for any $\varepsilon > 0$, we have

$$\kappa_n \triangleq \Pr\left(\left|\frac{\hat{K}}{n} - (1-\delta)\right| > \varepsilon\right) \to 0 \quad \text{as} \quad n \to \infty \tag{71}$$

We only consider the case $n(1-\delta-\varepsilon) \leq \hat{K} \leq n(1-\delta+\varepsilon)$ and declare error otherwise.

For any index set I of $\hat{K}$ elements , define $f(I)$ as the number of matching elements between $[n] \setminus I$ and $[n] \setminus I_{del}$.

For brevity let

$$d(I) \triangleq d_H(\mathscr{D}^{(1)}([n] \setminus I), \tilde{\mathscr{D}}_{\Phi}^{(2)}) \tag{72}$$

Now, the error probability of this algorithm is given as

$$P_e = \rho_n + \kappa_n + \Pr(\hat{I}_{del} \neq I_{del}) \tag{73}$$
$$P_e = \rho_n + \kappa_n + \Pr(\exists I \subseteq [n], |I| = K : d(I) \leq d(I_{del})) \tag{74}$$

Union bound yields

$$P_e \leq \rho_n + \kappa_n + \sum_{I \subseteq [n], |I| = \hat{K}} \Pr(d(I) \leq d(I_{del})) \tag{75}$$
$$= \rho_n + \kappa_n + \sum_{I \subseteq [n], |I| = \hat{K}} \Pr(d(I) - d(I_{del}) \leq 0) \tag{76}$$

Note that since the database entries are *i.i.d.*, $d(I) - d(I_{del})$ can be written as the difference of two Binomial random variables with parameters $(B(\hat{K} - f(I)), q_0(\Phi))$ and $(B(\hat{K} - f(I)), q_1(\Phi))$, respectively. Therefore we can use the Hoeffding's inequality [14] to obtain

$$\Pr(d(I) - d(I_{del}) \leq 0) \leq q^{B(\hat{K} - f_c(I))} \tag{77}$$

where

$$q \triangleq \exp\left(-\frac{1}{2}(q_0(\Phi) - q_1(\Phi))^2\right) < 1 \tag{78}$$

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

$$P_e \leq \rho_n + \kappa_n + \sum_{i=0}^{\hat{K}-1} \binom{n}{\hat{K}} q^{B(\hat{K}-i)} \tag{79}$$
$$= \rho_n + \kappa_n + \binom{n}{\hat{K}} \sum_{i=0}^{\hat{K}-1} q^{B(\hat{K}-i)} \tag{80}$$
$$= \rho_n + \kappa_n + \binom{n}{\hat{K}} \sum_{i=1}^{\hat{K}} q^{Bi} \tag{81}$$
$$= \rho_n + \kappa_n + \binom{n}{\hat{K}} q^B \sum_{i=0}^{\hat{K}-1} q^{Bi} \tag{82}$$
$$\leq \rho_n + \kappa_n + 2^{nH_b(\hat{K}/n)} q^B \frac{1-q^{B\hat{K}}}{1-q^B} \tag{83}$$
$$\leq \rho_n + \kappa_n + 2^{nH_b(\hat{K}/n)} q^B \frac{1}{1-q^B} \tag{84}$$
$$\leq \rho_n + \kappa_n + 2^{nH_b(\hat{K}/n)} q^B \frac{1}{1-q} \tag{85}$$
$$= \rho_n + \kappa_n + \frac{1}{1-q} 2^{nH_b(\hat{K}/n) - B\log\frac{1}{q}} \tag{86}$$

which vanishes as $n \to \infty$ if $B \geq \frac{nH_b(\hat{K}/n)}{\log\frac{1}{q}}$, which is satisfied for some $B = \Theta(n)$. Thus, a seed order $d = 1$ is sufficient for successful deletion detection. □