Database Matching Under Noisy Synchronization Errors

Serhat Bakirtas, Elza Erkip
Dept. of Electrical and Computer Engineering, New York University
{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, in the form of database matching, are successful in the presence of obfuscation. Motivated by 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 entry repetitions. By investigating different structures for the repetition pattern, replica detection and seeded deletion detection algorithms are devised and sufficient and necessary conditions for successful matching are derived. Finally, the impacts of some variations of the underlying assumptions, such as adversarial deletion model, seedless database matching and zero-rate regime, on the results are discussed. Overall, our results provide insights into the privacy-preserving publication of anonymized and obfuscated time-indexed data as well as the closely-related problem of the capacity of synchronization channels.

Index Terms

dataset, database, matching, de-anonymization, alignment, recovery, data, privacy, synchronization

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, there has been a growing concern over potential privacy leakage from anonymized data, approached from legal [1] and corporate [2] points of view. These concerns are also articulated in the respective literature through successful practical de-anonymization attacks on real data [3]–[17]. *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.

In the light of the above practical privacy attacks on databases, several groups initiated rigorous analyses of the graph matching problem [18]–[27]. Correlated graph matching has applications beyond privacy, such as image processing [28], computer vision [29], single-cell biological data alignment [30], [31] and DNA sequencing, which is shown to be equivalent to matching bipartite graphs [32]. Matching of correlated databases, also equivalent to bipartite graph matching, has also been investigated from information-theoretic [33]–[38] and statistical [39] perspectives. In [33], Cullina *et al.* introduced *cycle mutual information* as a correlation metric and derived sufficient conditions for successful matching and

This research was presented in part at the 2021 IEEE International Symposium on Information Theory (ISIT), the 2022 Asilomar Conference on Signals, Systems, and Computers and the 2022 IEEE Information Theory Workshop (ITW). It has also been in part submitted for conference publication. This work is supported in part by National Science Foundation grants 1815821 and 2148293, and NYU WIRELESS Industrial Affiliates.

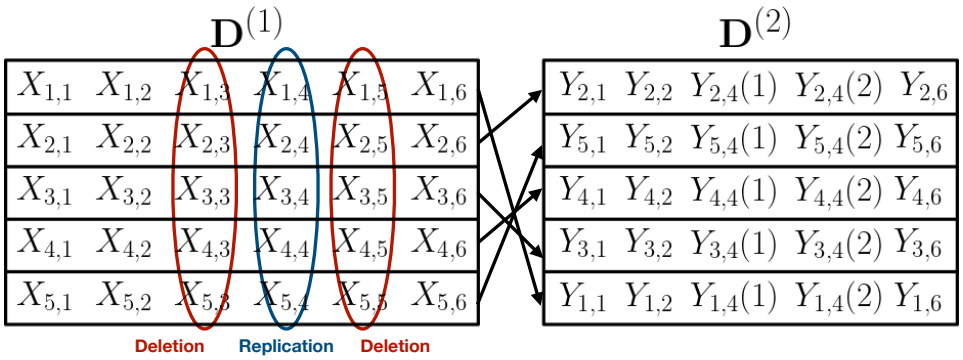

Fig. 1. An illustrative example of database matching under identical repetition, where each row experiences the same synchronization error. The columns circled in red are deleted whereas the fourth column, which is circled in blue, is repeated twice, *i.e.,* replicated. For each (i, j), $Y_{i,j}$ is 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 $\mathbf{D}^{(2)}$.

a converse result using perfect recovery as the error criterion. In [34], 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 necessary and sufficient conditions on the *database growth rate* for reliable matching, in the presence of noise on the database entries. In [35], Dai *et al.* considered the matching of a pair of databases with jointly Gaussian attributes with perfect recovery constraint. Similarly, in [39], Kunisky and Niles-Weed considered the same problem from the statistical perspective in different regimes of database size and under several recovery criteria. In [40], Kahraman and Nazer investigated the necessary and the sufficient conditions for detecting whether two Gaussian databases are correlated. More recently, motivated by the need for aligning single-cell data obtained from multiple biological sources/experiments [30], [31], in [41] Chen *et al.* investigated the matching of two noisy databases which are the noisy observations of a single underlying database under the fractional-error criterion, where the noise is assumed to be the Gaussian. They proposed a data-driven approach and analytically derived minimax lower bounds for successful matching.

Motivated by the synchronization errors in the sampling of time-series datasets, in this paper, we present a unified generalized framework of the database matching problem under noisy synchronization errors with near-exact recovery criterion. Specifically, we investigate the matching of Markov databases under arbitrary noise and synchronization errors. Our goal is to investigate necessary and sufficient conditions on the database growth rate [34] for the successful matching of database rows. The generalized Markov database model captures correlations of the attributes (columns), where synchronization errors, in the form of random entry deletions and replications, are followed by noise. As such, this paper generalizes the aforementioned work on database matching under only noise. Our setting is illustrated in Figure 1.

We consider two extreme regimes regarding the nature of synchronization errors, as results derived for these corner cases provides insights into the intermediate regime. To this end, first, we focus on the *identical repetition* setting where the repetition pattern is constant across rows. In other words, in this setting, deletions and replications only take place columnwise. We consider a two-phase matching scheme, where we first infer the underlying repetition structure by using permutation-invariant features of columns. This is followed by the matching phase which relies on the known replica and deletion locations. We show that as long as the databases are not independent, in the first phase replicas can be found with high probability through a series of hypothesis tests on the Hamming distances between columns. Furthermore, assuming *seed* rows whose identities are known in both databases [42], [43] we show that if the seed size Λ_n grows double-logarithmically with the number of rows m_n, where n denotes the column size, deletion locations can also be extracted. In the absence of noise, seeds are not needed and column histograms can be used to detect both replicas and deletions. Once the repetition (including deletions and replications)

locations are identified, in the second phase, we propose a joint typicality-based row matching scheme to derive sufficient conditions on the database growth rate for successful matching. Finally, we prove a tight converse result through a modified version of Fano's inequality, completely characterizing the matching capacity when the repetition pattern is constant across the rows.

Next, we focus on the other extreme, namely the *independent repetition* setting where the repetition pattern is independent in each row and there is no underlying repetition structure across rows. Under probabilistic side information on the deletion locations, we propose a row matching scheme and derive an achievable database growth rate. This, together with an outer bound obtained through Fano's inequality, provides upper and lower bounds on the matching capacity in the independent repetition setting. Comparing the bounds in the two extremes, we show that the matching capacity is lower and hence matching is more difficult under the independent repetition model. Finally, based on these two extreme models, we state bounds on the matching capacity for any intermediate repetition structure.

We also discuss the adversarial repetition model, where we assume that synchronization errors, in the form of column deletions, are chosen by a constrained adversary whose goal is to hinder the matching of databases, where the constraint is of the form of a fractional column deletion budget which naturally provides a trade-off between utility and privacy. Since this adversarial model forces us to focus on the worst-case scenario and in turn, prohibits the use of typicality and Fano's inequality, we propose an exact sequence matching and perform a more careful analysis of the worst case error, focusing on the Hamming distances between the rows (users) of the databases, as is the case in the adversarial channel literature [44], in our achievability and converse analyses. Under the identical repetition model, we completely characterize the adversarial matching capacity.

In addition to the characterization of the matching capacity under various assumptions, our results provide sufficient conditions on the number and the size for column histograms to be asymptotically unique. Since histograms naturally show up frequently in information theory, probability theory and statistics, this result could be of independent interest. In addition, our novel matching scheme in the independent repetition case can be directly converted to a decoding strategy for input-constrained noisy synchronization channels, a well-investigated model in the information theory literature [45]–[48].

A. Paper Organization

The organization of this paper is as follows: Section II contains the problem formulation. In Section III, our main results on the matching capacity under the identical repetition model are presented. Section IV contains our main results on the matching capacity under the independent repetition assumption. In Section V, we discuss the underlying model assumptions and investigate how variations on these assumptions impact some of the results. Finally, in Section VI the results and ongoing work are discussed.

B. Notations

In this paper we use the following notations:

- $[n]$ denotes the set of integers $\{1,...,n\}$.
- Matrices are denoted with uppercase bold letters. For a matrix $\mathbf{D}$, $D_{i,j}$ denotes the $(i,j)^{\text{th}}$ entry.
- a^n denotes a row vector consisting of scalars $a_1,\ldots,a_n$.
- Random variables are denoted by uppercase letters while their realizations are denoted by lowercase ones.
- The indicator of event E is denoted by $\mathbb{1}_E$.
- H and H_b denote the Shannon entropy and the binary entropy functions [49, Chapter 2], respectively.
- O, o, Θ, ω and Ω denote the standard asymptotic growth notations [50, Chapter 3].

- $D_{KL}(p_X \| q_X)$ denotes the Kullback-Leibler divergence [49, Chapter 2.3] between the probability distributions p_X and q_X. For scalars $p, q \in (0,1)$, $D(p\|q)$ denotes the Kullback-Leibler divergence between two Bernoulli distributions with respective parameters p and q. More formally,

$$D(p\|q) = (1-p)\log\frac{1-p}{1-q} + p\log\frac{p}{q} \tag{1}$$

- The logarithms, unless stated explicitly, are in base 2.

II. Problem Formulation

We use the following definitions, some of which are similar to [34], [36], [38], to formally describe our problem.

Definition 1. (Unlabeled Markov Database) An $(m_n, n, \mathbf{P})$ *unlabeled Markov database* is a randomly generated $m_n \times n$ matrix $\mathbf{D} = \{X_{i,j} \in \mathfrak{X} : i \in [m_n], j \in [n]\}$ whose rows are *i.i.d.* and follow a first-order stationary Markov process defined over the alphabet $\mathfrak{X} = \{1, \ldots, |\mathfrak{X}|\}$ with probability transition matrix $\mathbf{P}$ such that

$$\mathbf{P} = \gamma\mathbf{I} + (1-\gamma)\mathbf{U} \tag{2}$$

$$U_{i,j} = u_j > 0, \forall (i,j) \in \mathfrak{X}^2 \tag{3}$$

$$\sum_{j \in \mathfrak{X}} u_j = 1 \tag{4}$$

$$\gamma \in [0,1) \tag{5}$$

where $\mathbf{I}$ is the identity matrix. It is assumed that $X_{i,1} \overset{\text{i.i.d.}}{\sim} \pi = [u_1, \ldots, u_{|\mathfrak{X}|}]$, $i = 1, \ldots, m_n$, where π is the stationary distribution associated with $\mathbf{P}$.

Note that, the parameter γ determines the correlation among the columns of $\mathbf{D}^{(1)}$. Specifically, $\gamma = 0$ corresponds to the case where $X_{i,j}$ are *i.i.d.*

Definition 2. (Repetition Matrix) The *repetition matrix* $\mathbf{S}$ is a random matrix of size $m_n \times n$ with the following structure:

- $\mathbf{S}$ consists of independent mutually exclusive blocks of $W_n = \Theta(n^{d_{\text{rep}}})$ consecutive rows.
- Each block of size $W_n \times n$ is obtained by repeating a row vector W_n times, where the row vector consists of n *i.i.d.* entries drawn from a discrete probability distribution p_S with a finite integer support $\{0, \ldots, s_{\max}\}$.

Here W_n and d_{rep} are called *repetition block size* and *repetition order*, respectively. Furthermore, the parameter $\delta \triangleq p_S(0)$ is called the *deletion probability*.

In most of the paper, we assume a random repetition pattern as in Definition 2. In Section V-A, we will discuss the effects of adversarial worst-case repetition patterns.

Definition 3. (Correlated Repeated Database, Labeling Function) Let $\mathbf{D}^{(1)}$ be an (m_n, n, P) unlabeled Markov database, $\mathbf{S}$ be the repetition matrix, Θ_n be a uniform permutation of $[m_n]$ with $\mathbf{D}^{(1)}$, $\mathbf{S}$ and Θ_n independently chosen as in Definitions 1 and 2. Also, let $p_{Y|X}$ be a conditional probability distribution with both X and Y taking values from $\mathfrak{X}$. Given $\mathbf{D}^{(1)}$, $\mathbf{S}$ and $p_{Y|X}$, the pair $(\mathbf{D}^{(2)}, \Theta_n)$ is called the *labeled repeated database* if the $(i,j)^{\text{th}}$ entry $X_{i,j}$ of $\mathbf{D}^{(1)}$ and the $(\Theta_n(i), j)^{\text{th}}$ entry $Y_{\Theta_n(i),j}$ of $\mathbf{D}^{(2)}$ have the following relation:

$$Y_{\Theta_n(i),j} = \begin{cases} E, & \text{if } S_{\Theta_n(i),j} = 0 \\ Y^{S_{\Theta_n(i),j}} & \text{if } S_{\Theta_n(i),j} \geq 1 \end{cases} \tag{6}$$

for $(i,j) \in [m_n] \times [n]$, where $Y^{S_{\Theta_n(i),j}}$ is a random row vector of length $S_{\Theta_n(i),j}$ with the following probability distribution, conditioned on $X_{i,j}$

$$\Pr\left(Y^{S_{\Theta_n(i),j}} = y^{S_{\Theta_n(i),j}} \middle| X_{i,j} = x\right) = \prod_{l=1}^{S_{\Theta_n(i),j}} p_{Y|X}(y_l|x) \tag{7}$$

and $Y_{\Theta_n(i),j} = E$ corresponds to $Y_{\Theta_n(i),j}$ being the empty string. Θ_n and $\mathbf{D}^{(2)}$ are called the *labeling function* and *correlated repeated database*, respectively.

Note that $S_{\Theta_n(i),j}$ indicates the times $X_{i,j}$ is repeated (including deletions and replications). When $S_{\Theta_n(i),j} = 0$, $X_{i,j}$ is said to be *deleted* (repeated zero times) and when $S_{\Theta_n(i),j} > 1$, $X_{i,j}$ is said to be *replicated* $S_{\Theta_n(i),j}$ times (repeated $S_{\Theta_n(i),j}$ times).

The respective rows $X_{i_1}^n$ and $Y_{i_2}^{K_n}$ of $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$ are said to be *matching rows*, if $\Theta_n(i_1) = i_2$, where $K_n \triangleq \sum_{j=1}^n S_{i_2,j}$.

In our model, the correlated repeated database $\mathbf{D}^{(2)}$ is obtained by permuting the rows of the unlabeled Markov database $\mathbf{D}^{(1)}$ with the uniform permutation Θ_n followed by repetition based on the repetition matrix $\mathbf{S}$ and introduction of noise through $p_{Y|X}$. The relationship between $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$, as described in Definition 3, is illustrated in Figure 2. As we formalize later, the goal is to recover the labeling function Θ_n based on the observations of $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$.

Note that (7) states that we can treat $Y_{\Theta_n(i),j}$ as the output of the discrete memoryless channel (DMC) $p_{Y|X}$ with input sequence consisting of $S_{\Theta_n(i),j}$ copies of $X_{i,j}$ concatenated together. We stress that $p_{Y|X}$ is a general model, capturing any distortion and noise on the database entries, though we refer to this as "noise" in this paper.

We are mainly interested in two extremes for the repetition block size W_n:

- Every row of $\mathbf{D}^{(1)}$ experiences the same repetition which we call *identical repetition*. More formally, $\mathbf{S}$ is a matrix consisting of m_n identical copies of a row vector S^n called the *column repetition pattern* and $W_n = m_n \sim 2^{nR}$ which we denote by $d_{\text{rep}} = \infty$.
- Rows of $\mathbf{D}^{(1)}$ experience *i.i.d.* repetition which we call *independent repetition*. More formally, $\mathbf{S}$ has *i.i.d.* rows and $W_n = 1$ which we denote by $d_{\text{rep}} = 0$.

We will observe that these two models pose different challenges to matching and in turn necessitate different solutions with different implications. After focusing on the two extremes described above in Sections III and IV, we will discuss the intermediate regimes for the repetition block size W_n in Sections IV-A and IV-B.

As discussed in Sections III and IV, inferring the repetition pattern, particularly deletions, is a difficult, if not impossible, task. Therefore, we assume the availability of *seeds* to help with the inference of the underlying repetition pattern, similar to database matching [36] and graph matching [42], [43] settings.

Definition 4. (Seeds) A *seed* is a pair of matching rows whose labels are known universally. A *batch of* Λ_n *seeds* $(\mathbf{G}^{(1)}, \mathbf{G}^{(2)})$ is a batch of Λ_n correctly-matched row pairs. Λ_n is called the *seed size*.

The relation between the repetition patterns of the database pair $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$ and the seeds $(\mathbf{G}^{(1)}, \mathbf{G}^{(2)})$ will be clarified in the relevant sections. Furthermore, note that in Definition 4, for convenience the seeds are assumed to be additional to the databases.

Throughout Sections III and IV, we assume a double logarithmic seed size $\Lambda_n = \Omega(\log\log m_n)$. We will discuss the effects of not having seeds in Section V-B.

Besides the seeds, we assume that the locations of some deleted entries are revealed. This is formalized in the following definition:

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

Definition 5. (Partial Deletion Location Information) Given the repetition matrix $\mathbf{S}$ with the repetition block size W_n, the *partial deletion location information* $\mathbf{A}$ is an $m_n \times n$ random matrix, with the following conditional distribution on $\mathbf{S}$ and its structure:

$$\Pr(A_{iW_n+1,j} = 1|\mathbf{S}) = \alpha \mathbb{1}_{[S_{iW_n+1,j}=0]} \tag{8}$$

$$A_{iW_n+l,j} = A_{iW_n+1,j} \tag{9}$$

$$\forall i \in \left\{0,\ldots,\left\lfloor \frac{m_n}{W_n} \right\rfloor\right\}, \forall j \in [n], \forall l \in [W_n - 1] \tag{10}$$

where $A_{i,j} = 1$ corresponds to $\mathbf{D}^{(1)}_{\Theta_n(i),j}$ being revealed as deleted and $A_{i,j} = 0$ corresponds to either $\mathbf{D}^{(1)}_{\Theta_n(i),j}$ not being deleted or not being revealed after deletion. The parameter $\alpha \in [0,1]$ is called the *deletion detection probability*.

Definition 5 states that in a given $W_n \times n$ block in which the repetition matrix has identical rows, the location of each deleted column is revealed with probability α. Since the columns of $\mathbf{S}$ are i.i.d. in a given such block and $\mathbf{S}$ and $\mathbf{D}^{(1)}$ are independent, each deleted column is revealed independently of the other columns of $\mathbf{S}$ and $\mathbf{D}^{(1)}$. Furthermore, since for any $i_1 \neq i_2$, $\forall j_1, j_2 \in [n]$, $\forall l_1, l_2 \in [W_n]$, $S_{i_1W_n+l_1,j_1}$ and $S_{i_2W_n+l_2,j_1}$ are independent, leading to the independence of $A_{i_1W_n+l_1,j_1}$ and $A_{i_2W+l_2,j_1}$ $i_1 \neq i_2$, $\forall j_1, j_2 \in [n]$, $\forall l_1, l_2 \in [W_n]$. In other words, any two entries of $\mathbf{A}$ located in different columns and/or different such $W_n \times n$ blocks are independent.

Definition 6. (Successful Matching Scheme) A *matching scheme* is a sequence of mappings $\phi_n : (\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \mathbf{A}) \mapsto \hat{\Theta}_n$ where $\mathbf{D}^{(1)}$ is the unlabeled Markov database, $\mathbf{D}^{(2)}$ is the correlated column repeated database, $(\mathbf{G}^{(1)}, \mathbf{G}^{(2)})$ are seeds, $\mathbf{A}$ is the partial deletion location information and $\hat{\Theta}_n$ is the estimate of the correct labeling function Θ_n. The scheme ϕ_n is *successful* if

$$\Pr\left(\hat{\Theta}_n(J) \neq \Theta_n(J)\right) \to 0 \text{ as } n \to \infty \tag{11}$$

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

We stress that both in database matching and correlation detection settings, the relationship between the row size m_n, the column size n and the database distribution parameters are the parameters of interest [39], [40], [51]. Note that for fixed column size n, as the row size m_n increases, matching becomes harder. This is because 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 candidate row set. Furthermore, as stated in [39, Theorem 1.2], for distributions with parameters constant in n and m_n, the regime of interest is the logarithmic regime where $n \sim \log m_n$. Thus, we utilize the *database growth rate* introduced in [34] to characterize the relationship between the row size m_n and the column size n.

Definition 7. (Database Growth Rate) The *database growth rate* R of an unlabeled Markov database with m_n rows and n columns is defined as

$$R = \lim_{n\to\infty} \frac{1}{n} \log m_n \tag{12}$$

In Sections III and IV, we assume that the database growth rate R is positive. We will discuss the zero-rate regime $R = 0$ in Section V-C.

Definition 8. (Achievable Database Growth Rate) Consider a sequence of $(m_n, n, \mathbf{P})$ unlabeled Markov databases, a repetition probability distribution p_S, a repetition order d_{rep}, a noise distribution $p_{Y|X}$ and the resulting sequence of correlated repeated databases. For a seed size Λ_n and a deletion detection probability α, 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 9. (Matching Capacity) The *matching capacity* $C(d_{\text{rep}}, \alpha)$ is the supremum of the set of all achievable rates corresponding to a probability transition matrix $\mathbf{P}$, repetition probability distribution p_S, repetition order d_{rep}, noise distribution $p_{Y|X}$, seed size Λ_n and a deletion detection probability α.

In this paper, our goal is to characterize the matching capacity $C(d_{\text{rep}}, \alpha)$ in different regimes of the parameters by providing database matching schemes as well as upper bounds on all achievable database growth rates.

III. Matching Capacity For Identical Repetition

In this section, we present the matching capacity $C(d_{\text{rep}}, \alpha)$ for an identical repetition pattern ($W_n = m_n$, $d_{\text{rep}} = \infty$) with seed size $\Lambda_n = \Omega(\log\log m_n)$. We will show that when $\Lambda_n = \Omega(\log\log m_n)$, the repetition pattern, including the deletion locations, can be inferred. Therefore partial deletion location information $\mathbf{A}$ will become obsolete in this case and our results hold for any $\alpha \geq 0$.

We state the main result of this section in 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)* row matching. Then, we prove the converse part. Finally, we focus on the noiseless setting as a special case where we prove that we can devise a new detection algorithm specific to the noiseless model which renders the seeds obsolete.

Theorem 1. ***(Matching Capacity for Identical Repetition)*** *Consider a probability transition matrix $\mathbf{P}$, a column repetition distribution p_S with an identical repetition pattern and a noise distribution $p_{Y|X}$. Then, for any deletion detection probability $\alpha \geq 0$ and a seed size $\Lambda_n = \Omega(\log\log m_n)$, the matching capacity is*

$$C(\infty, \alpha) = \lim_{n\to\infty} \frac{I(X^n; Y^{K_n}, S^n)}{n} \tag{13}$$

where X^n is a Markov chain with probability transition matrix $\mathbf{P}$ and stationary distribution μ, $S_i \overset{iid}{\sim} p_S$ and $Y^{K_n} = Y_1^{S_1}, \ldots, Y_n^{S_n}$ with $K_n = \sum_{j=1}^{n} S_j$ such that

$$\Pr(Y_i^{S_i} = y^{S_i} | X_i = x_i) = \begin{cases} \prod_{j=1}^{S_i} p_{Y|X}(y_j|x_i) & \text{if } S_i > 0 \\ \mathbb{1}_{[y^{s_i} = E]} & \text{if } S_i = 0 \end{cases}, \tag{14}$$

for all $i \in [n]$ with E denoting the empty string.

Because of the independence of X^n and S^n, (13) can also be represented as

$$C(\infty, \alpha) = \lim_{n\to\infty} \frac{I(X^n; Y^{K_n} | S^n)}{n}. \tag{15}$$

Hence, Theorem 1 states that although the repetition pattern S^n is not known a-priori, for a seed size $\Lambda_n = \Omega(\log\log m_n)$, we can achieve a database growth rate as if we knew S^n. Since the utility of seeds increases with the seed size Λ_n, we will focus on $\Lambda_n = \Theta(\log\log m_n)$, which we show is sufficient to achieve the matching capacity.

The rest of this section is on the proof of Theorem 1. In Section III-A, we discuss our noisy replica detection algorithm which does not utilize the seeds 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 proposing a typicality-based matching scheme for rows, which is performed once replicas and deletions are detected. In Section III-D, we prove the converse part of Theorem 1. Finally, in Section III-E, we focus on the special case of no noise on the repeated entries and provide a single repetition (replica and deletion) detection algorithm which does not require any seeds.

Note that when the two databases are independent, Theorem 1 states that the matching capacity becomes zero, hence our results trivially hold. As a result, throughout this section, we assume that the two databases are not independent. Furthermore, our achievability result assumes $\alpha = 0$ and its proof holds for any $\alpha \geq 0$.

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 replicas, not deletions. Note that our replica detection algorithm does not require any seeds unlike seeded deletion detection discussed in Section III-B.

Our proposed replica detection algorithm adopts the *Hamming distance between consecutive columns* of $\mathbf{D}^{(2)}$ as a permutation-invariant feature of the columns. The permutation-invariance allows us to perform replica detection on $\mathbf{D}^{(2)}$ with no a-priori information on Θ_n.

Let K_n denote the number of columns of $\mathbf{D}^{(2)}$, $C_j^{m_n}$ denote the j^{th} column of $\mathbf{D}^{(2)}$, $j = 1, \ldots, K_n$. The replica detection algorithm works as follows: We first compute the Hamming distances $d_H(C_j^{m_n}, C_{j+1}^{m_n})$ between consecutive columns $C_j^{m_n}$ and $C_{j+1}^{m_n}$, for $j \in [K_n - 1]$. For some average Hamming distance threshold $\tau \in (0,1)$ chosen based on $\mathbf{P}$ and $p_{Y|X}$ (See Appendix A), 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 correspond to distinct columns of $\mathbf{D}^{(1)}$ 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 replica relationship between the columns $C_j^{m_n}$ and $C_{j+1}^{m_n}$ of $\mathbf{D}^{(2)}$, $j = 1, \ldots, K_n - 1$. The total probability of replica detection error diminishes as $n \to \infty$, that is*

$$\Pr\Big(\bigcup_{j=1}^{K_n-1} E_j\Big) \to 0 \text{ as } n \to \infty. \tag{16}$$

Proof. See Appendix A. □

B. Deletion Detection Using Seeds

Since the replica detection algorithm discussed in Section III-A only uses $\mathbf{D}^{(2)}$ and thus only the retained columns, it cannot detect column deletions. We next propose a deletion detection algorithm which uses seeds.

Let $(\mathbf{G}^{(1)}, \mathbf{G}^{(2)})$ be a batch of $\Lambda_n = \Theta(\log\log m_n)$ seeds with the identical repetition pattern S^n. In other words, let $(\mathbf{G}^{(1)}, \mathbf{G}^{(2)})$ have the same repetition pattern as $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$. 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 $\hat{K}_n$. At this step, we only have deletions.

Next, for each index pair $(i, j) \in [n] \times [\hat{K}_n]$, we compute the Hamming distance $d_H(C_i^{(1)}, C_j^{(2)})$ between the i^{th} column $C_i^{(1)}$ of $\mathbf{G}^{(1)}$ and the j^{th} column $C_j^{(2)}$ of $\tilde{\mathbf{G}}^{(2)}$. More formally, we compute

$$d_H(C_i^{(1)}, C_j^{(2)}) = \sum_{t=1}^{\Lambda_n} \mathbb{1}_{\left[G_{t,i}^{(1)} \neq \tilde{G}_{t,j}^{(2)}\right]}. \tag{17}$$

Then, for each index $i \in [n]$, the algorithm decides $C_i^{(1)}$ is retained (not deleted) only if there exists a column $C_j^{(2)}$ in $\tilde{\mathbf{G}}^{(2)}$ with $d_H(C_i^{(1)}, C_j^{(2)}) \leq \Lambda_n\bar{\tau}$, for some average Hamming distance threshold $\bar{\tau} \in (0,1)$ chosen based on $\mathbf{P}$ and $p_{Y|X}$. In this case, we assign $\hat{I}_i = 0$. Otherwise, the algorithm decides $C_i^{(1)}$ is

deleted, assigning $\hat{I}_i = 1$. At the end of this procedure, the algorithm outputs an estimate $\hat{I}^n = (\hat{I}_1, \ldots, \hat{I}_n)$ of the true deletion pattern $I^n_{\text{del}} = (I_1, \ldots, I_n)$. Here, for each $i \in [n]$ we have

$$I_i \triangleq \mathbb{1}_{[S_i=0]} \tag{18}$$

$$\hat{I}_i \triangleq \mathbb{1}_{\left[\exists j \in [\hat{K}_n] : d_H(C_i^{(1)}, C_j^{(2)}) \leq \Lambda_n \bar{\tau}\right]} \tag{19}$$

Note that such a Hamming distance-based strategy depends on pairs of matching entries in a pair of seed rows in $\mathbf{G}^{(1)}$ and $\tilde{\mathbf{G}}^{(2)}$ having a higher probability of being equal than non-matching entries. More formally, WLOG, let $S_j \neq 0$ and $\tilde{X}_{i,j}$ and $\tilde{Y}_{i,j}$ denote the respective $(i,j)^{\text{th}}$ entries of $\mathbf{G}^{(1)}$ and $\tilde{\mathbf{G}}^{(2)}$. Given a matching pair $(\tilde{X}_{i,j}, \tilde{Y}_{i,j})$ of entries and any non-matching pair $(\tilde{X}_{i,l}, \tilde{Y}_{i,j})$, $l \neq j$ we need

$$\Pr(\tilde{Y}_{i,j} \neq \tilde{X}_{i,j}) < \Pr(\tilde{Y}_{i,j} \neq \tilde{X}_{i,l}) \tag{20}$$

which may not be true in general.

For example, suppose we have a binary uniform *i.i.d.* distribution, *i.e.,* $\mathfrak{X} = \{0,1\}$ with $\gamma = 0$ and $u_1 = 1/2$ (recall Definition 1). Further assume that $p_{Y|X}$ follows BSC(q), *i.e.* $p_{Y|X}(x|x) = 1-q$, $x = 0,1$. Note that when $q > 1/2$, equation (20) is not satisfied. However, in this example, we can flip the labels in Y by applying the bijective remapping $\sigma = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ to Y in order to satisfy equation (20).

Thus, as long as such a permutation σ of $\mathfrak{X}$ satisfying equation (20) 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 follows the above steps computing $d_H(C_i^{(1)}, C_j^{(2)}(\sigma))$ for each index pair $(i,j) \in [n] \times [\hat{K}_n]$ and outputs the deletion pattern estimate $\hat{I}^n(\sigma) = (\hat{I}_1(\sigma), \ldots, \hat{I}_n(\sigma))$ where

$$\hat{I}_i(\sigma) \triangleq \mathbb{1}_{\left[\exists j \in [\hat{K}_n] : d_H(C_i^{(1)}, C_j^{(2)}(\sigma)) \leq \Lambda_n \bar{\tau}\right]} \tag{21}$$

and $C_j^{(2)}(\sigma)$ is the j^{th} column of $\tilde{\mathbf{G}}^{(2)}_\sigma$.

The following lemma states that such a bijective mapping σ always exists and for a seed size $\Lambda_n = \Theta(\log n) = \Theta(\log \log m_n)$, 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 such that equation* (20) *holds after the remapping. In addition, for a seed size $\Lambda_n = \Theta(\log n) = \Theta(\log \log m_n)$, using the algorithm above, we have*

$$\Pr\left(\hat{I}(\sigma) = I_{del}\right) \to 1 \text{ as } n \to \infty. \tag{22}$$

Proof. See Appendix B. □

C. Row Matching Scheme and Achievability

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

Proof of Achievability of Theorem 1. Let S^n be the underlying column repetition pattern and $K_n \triangleq \sum_{j=1}^n S_j$ 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 in 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 $\Theta_n \circ \tilde{\mathbf{D}}^{(1)}$ if $S_j \neq 0$, and an erasure column otherwise.

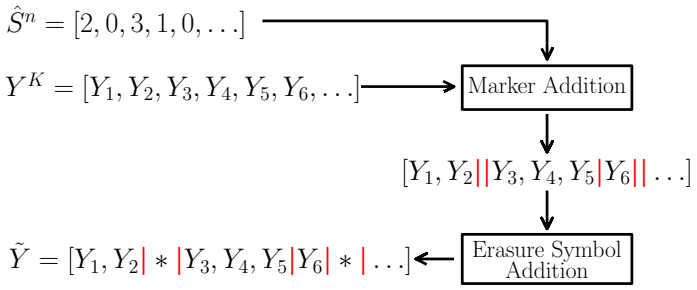

Fig. 3. An example of the construction of $\tilde{\mathbf{D}}^{(2)}$, as described in Step 3 of the proof of Theorem 1 in Section III-C, illustrated over a pair of rows X^n of $\mathbf{D}^{(1)}$ and Y^K of $\mathbf{D}^{(2)}$. After these steps, in Step 4 we check the joint typicality of the rows X^n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

4) Fix $\varepsilon > 0$. Match the l^{th} row $Y_l^{K_n}$ of $\tilde{\mathbf{D}}^{(2)}$ with the i^{th} row X_i^n of $\tilde{\mathbf{D}}^{(1)}$ if X_i^n is the only row of $\tilde{\mathbf{D}}^{(1)}$ jointly ε-typical with $Y_l^{K_n}$ according to p_{X^n,Y^{K_n},S^n}, where $S_i \overset{\text{iid}}{\sim} p_S$ and $Y^{K_n} = Y_1^{S_1}, \ldots, Y_n^{S_n}$ such that

$$p_{X^n,Y^K|S^n}(x^n, y^k|s^n) = p_{X^n}(x^n) \prod_{i:s_i>0} \left(\prod_{j=1}^{s_i} p_{Y|X}((y^{s_i})_j|x_i) \right) \prod_{i:s_i=0} \mathbb{1}_{[y^{s_i}=*]} \tag{23}$$

with $y^k = y^{s_1} \ldots y^{s_n}$. Assign $\hat{\Theta}_n(i) = l$. If there is no such jointly typical row, or there are more than one, declare an error.

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

Using the union bound and the generalized Asymptotic Equipartition Property (AEP) [34, Proposition 3], the total probability of error of this scheme (as in (11)) can be bounded as follows

$$P_e \leq 2^{nR} 2^{-n(\bar{I}(X;Y^S,S)-3\varepsilon)} + \varepsilon + \rho_n + \mu_n \tag{24}$$

where $\bar{I}(X;Y^S,S)$ is the mutual information rate [52] defined as

$$\bar{I}(X;Y^S,S) \triangleq \lim_{n\to\infty} \frac{1}{n} I(X^n;Y^{K_n},S^n) \tag{25}$$

Note that since m_n is exponential in n, from Lemma 1 we have $\rho_n \to 0$. Furthermore, since $\Lambda_n = \Theta(\log n)$, from Lemma 2 we have $\mu_n \to 0$ as $n \to \infty$. Thus $P_e \leq \varepsilon$ as $n \to \infty$ if

$$R < \lim_{n\to\infty} \frac{1}{n} I(X^n;Y^{K_n},S^n) \tag{26}$$

concluding the proof of the achievability part. □

D. Converse

In this subsection, we prove that the database growth rate achieved in Theorem 1 is in fact tight using a genie-aided proof where the column repetition pattern S^n is known. 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 when S^n is given. Thus, the genie-aided proof holds for any seed size Λ_n.

Proof of Converse of Theorem 1. While Theorem 1 is stated for $\Lambda_n = \Omega(\log\log m_n)$, in the converse we assume any seed size Λ_n. We prove the converse using the modified Fano's inequality presented in [34]. 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,

$$P_e \triangleq \Pr\left(\Theta_n(J) \neq \hat{\Theta}_n(J)\right), \quad J \sim \text{Unif}([m_n]) \tag{27}$$

Suppose $P_e \to 0$ as $n \to \infty$. Furthermore, let S^n be the repetition pattern and $K_n = \sum_{j=1}^{n} S_j$. Since Θ_n is a uniform permutation, from Fano's inequality, we have

$$H(\Theta_n) \leq 1 + m_n P_e \log m_n + I(\Theta_n; \mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \mathbf{A}, S^n) \tag{28}$$

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

$$\begin{aligned}
I(\Theta_n; \mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \mathbf{A}, S^n) &= I(\Theta_n; \mathbf{D}^{(1)} | \mathbf{D}^{(2)}, \mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \mathbf{A}, S^n) && (29)\\
&\leq I(\Theta_n, \mathbf{D}^{(2)}, \mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \mathbf{A}, S^n; \mathbf{D}^{(1)}) && (30)\\
&\leq I(\Theta_n, \mathbf{D}^{(2)}, S^n; \mathbf{D}^{(1)}) && (31)\\
&= I(\Theta_n, \mathbf{D}^{(2)}; \mathbf{D}^{(1)} | S^n) && (32)\\
&= \sum_{i=1}^{m_n} I(X_i^n; Y_{\Theta_n(i)}^{K_n} | S_{\Theta_n(i)}^n) && (33)\\
&= m_n I(X^n; Y^{K_n} | S^n) && (34)\\
&= m_n I(X^n; Y^{K_n}, S^n) && (35)
\end{aligned}$$

where (31) follows from the fact that given the repetition pattern S^n, the seeds $(\mathbf{G}^{(1)}, \mathbf{G}^{(2)})$ and $\mathbf{A}$ do not offer any additional information on Θ_n. Equation (33) follows from the conditional independence of the non-matching rows given S^n. Equation (34) follows from the fact that the matching rows are identically distributed conditioned on the repetition pattern S^n. Finally, (35) follows from the independence of X^n and S^n.

Note that from Stirling's approximation [50, Chapter 3.2] and the uniformity of Θ_n, we get

$$\begin{aligned}
H(\Theta_n) &= \log m_n! && (36)\\
&= m_n \log m_n - m_n \log e + O(\log m_n) && (37)
\end{aligned}$$

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{m_n n} H(\Theta_n) &= \lim_{n\to\infty} \frac{1}{m_n n} \left[m_n \log m_n - m_n \log e + O(\log m_n) \right] && (38)\\
&= \lim_{n\to\infty} \frac{1}{n} \log m_n && (39)\\
&= R && (40)
\end{aligned}$$

Finally, from (28)-(40) we obtain

$$\begin{aligned}
R &= \lim_{n\to\infty} \frac{1}{m_n n} H(\Theta_n) && (41)\\
&\leq \lim_{n\to\infty} \left[\frac{1}{m_n n} + P_e R + \frac{1}{n} I(X^n; Y^{K_n}, S^n) \right] && (42)\\
&= \lim_{n\to\infty} \frac{I(X^n; Y^{K_n}, S^n)}{n} && (43)
\end{aligned}$$

where (43) follows from the fact that $P_e \to 0$ as $n \to \infty$. □

E. Noiseless Setting

Lemmas 1 and 2 state that given a seed size Λ_n double logarithmic with the row size m_n, the repetition pattern can be inferred through the aforementioned replica and deletion detection algorithms for any noise distribution $p_{Y|X}$. Thus, the results of Section III-A through Section III-C trivially apply to the noiseless setting where

$$p_{Y|X}(y|x) = \mathbb{1}_{[y=x]} \, \forall (x, y) \in \mathfrak{X}^2. \tag{44}$$

We note that when there is no noise, the capacity expression of Theorem 1 (Equation 13) can be further simplified as

$$C(\infty,0) = (1-\delta)^2 \sum_{r=0}^{\infty} \delta^r H(X_0|X_{-r-1}). \tag{45}$$

In this subsection, we show that in the noiseless setting, seeds can be made obsolete by the use of a novel detection algorithm. In other words, in the noiseless setting, we show that Theorem 1 can be extended to any seed size Λ_n.

Theorem 2. ***(Noiseless Matching Capacity for Identical Repetition)*** *Consider a probability transition matrix* $\mathbf{P}$ *and a repetition probability distribution* p_S*. Suppose there is no noise,* i.e.,

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

Then, the matching capacity is

$$C(\infty,\alpha) = (1-\delta)^2 \sum_{r=0}^{\infty} \delta^r H(X_0|X_{-r-1}) \tag{47}$$

for any seed size Λ_n *and deletion detection probability* $\alpha \geq 0$*. Here* $\delta \triangleq p_S(0)$ *is the deletion probability and* $H(X_0|X_{-r-1})$ *is the entropy rate associated with the probability transition matrix*

$$\mathbf{P}^{r+1} = \gamma^{r+1}\mathbf{I} + (1-\gamma^{r+1})\mathbf{U} \tag{48}$$

The capacity can further be simplified as

$$C(\infty,\alpha) = \frac{(1-\delta)(1-\gamma)}{(1-\gamma\delta)}[H(\pi) + \sum_{i\in\mathfrak{X}} u_i^2 \log u_i] - (1-\delta)^2 \sum_{r=0}^{\infty} \delta^r \sum_{i\in\mathfrak{X}} u_i \eta_{r,i} \log \eta_{r,i} \tag{49}$$

where

$$\eta_{r,i} \triangleq (1-u_i)\gamma^{r+1} + u_i. \tag{50}$$

Observe that the RHS of (47) is the mutual information rate for an erasure channel with erasure probability δ with first-order Markov ($\mathbf{P}$) inputs, as stated in [53, Corollary II.2]. Thus, Theorem 2 states that we can achieve the erasure bound which assumes a-priori knowledge of the column repetition pattern.

The proof of Theorem 2 hinges on the observation that in the noiseless setting deletion and replica detection can be performed without seeds. Inspired by the idea of extracting permutation-invariant features as done in Section III-A, our noiseless repetition detection algorithm uses the histogram (and equivalently the type) of each column of $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$ as the permutation-invariant feature. Our repetition detection algorithm works as follows: First, for tractability, we "collapse" the Markov chain into a binary-valued one. We pick a symbol x from the alphabet $\mathfrak{X}$, WLOG $x = 1$, and define the *collapsed* databases $\tilde{\mathbf{D}}^{(1)}$ and $\tilde{\mathbf{D}}^{(2)}$ as follows:

$$\tilde{\mathbf{D}}_{i,j}^{(r)} = \begin{cases} 1 & \text{if } \mathbf{D}_{i,j}^{(r)} = 1 \\ 2 & \text{if } \mathbf{D}_{i,j}^{(r)} \neq 1 \end{cases}, \forall (i,j),\, r = 1,2 \tag{51}$$

Next, we construct the collapsed histogram vectors $\tilde{H}^{(1),n}$ and $\tilde{H}^{(2),K_n}$ as

$$\tilde{H}_j^{(r)} = \sum_{i=1}^{m_n} \mathbb{1}_{\left[\tilde{D}_{i,j}^{(r)}=2\right]}, \quad \begin{cases} \forall j \in [n], & \text{if } r = 1 \\ \forall j \in [K_n] & \text{if } r = 2 \end{cases} \tag{52}$$

Then, the algorithm declares the j^{th} column deleted if $\tilde{H}_j^{(1)}$ is absent in $\tilde{H}^{(2),K_n}$ and declares the j^{th} column replicated s times if $\tilde{H}_j^{(1)}$ is present $s \geq 1$ times in $\tilde{H}^{(2),K_n}$.

Note that as long as column histograms $\tilde{H}_j^{(1)}$ of the collapsed database $\tilde{\mathbf{D}}^{(1)}$ are unique, this detection process is error-free.

The following lemma provides conditions for the asymptotic uniqueness of column histograms $\tilde{H}_j^{(1)}$, $j \in [n]$.

Lemma 3. ***(Asymptotic Uniqueness of the Column Histograms)*** *Let $\tilde{H}_j^{(1)}$ denote the histogram of the j^{th} column of $\tilde{\mathbf{D}}^{(1)}$, as in* (52)*. Then, for $m_n = \omega(n^4)$, we have*

$$\Pr\left(\exists i, j \in [n], i \neq j, \tilde{H}_i^{(1)} = \tilde{H}_j^{(1)}\right) \to 0 \text{ as } n \to \infty. \tag{53}$$

Proof. See Appendix C. □

When the databases are not collapsed, the order relation given in Lemma 3 can be tightened. See Section V-C for more details.

Note that by Definition 7, the row size m_n is exponential in the column size n and the order relation of Lemma 3 is automatically satisfied.

Next, we present the proof of the achievability part of Theorem 2.

Proof of Achievability of Theorem 2. Let S^n be the underlying repetition pattern and $K_n \triangleq \sum_{j=1}^n S_j$ be the number of columns in $\mathbf{D}^{(2)}$. Our matching scheme consists of the following steps:

1) Construct the collapsed histogram vectors $\tilde{H}^{(1),n}$ and $\tilde{H}^{(2),K_n}$ as in (52).

2) Check the uniqueness of the entries $\tilde{H}_j^{(1)}$ $j \in [n]$ of $\tilde{H}^{(1),n}$. If there are at least two which are identical, declare a *detection error* whose probability is denoted by μ_n. Otherwise, proceed with Step 3.

3) If $\tilde{H}_j^{(1)}$ is absent in $\tilde{H}^{(2),K_n}$, declare it deleted, assigning $\hat{S}_j = 0$. Note that, conditioned on the uniqueness of the column histograms $\tilde{H}_j^{(1)}$ $\forall j \in [n]$, this step is error-free.

4) If $\tilde{H}_j^{(1)}$ is present $s \geq 1$ times in $\tilde{H}^{(2),K_n}$, assign $\hat{S}_j = s$. Again, if there is no detection error in Step 2, this step is error-free. Note that at the end of this step, provided there are no detection errors, we recover S^n, *i.e.*, $\hat{S}^n = S^n$.

5) Based on $\hat{S}^n$, $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$, construct $\bar{\mathbf{D}}^{(2)}$ as the following:
- If $\hat{S}_j = 0$, the j^{th} column of $\bar{\mathbf{D}}^{(2)}$ is a column consisting of erasure symbol $* \notin \mathfrak{X}$.
- If $\hat{S}_j \geq 1$, the j^{th} column of $\bar{\mathbf{D}}^{(2)}$ is the j^{th} column of $\mathbf{D}^{(1)}$.

Note that after the removal of the additional replicas and the introduction of the erasure symbols, $\bar{\mathbf{D}}^{(2)}$ has n columns.

6) Fix $\varepsilon > 0$. Let $q_{\bar{Y}|X}$ be the probability transition matrix of an erasure channel with erasure probability δ, that is $\forall (x, \bar{y}) \in \mathfrak{X} \times (\mathfrak{X} \cup \{*\})$

$$q_{\bar{Y}|X}(\bar{y}|x) = \begin{cases} 1-\delta & \text{if } \bar{y} = x \\ \delta & \text{if } \bar{y} = * \end{cases}. \tag{54}$$

We consider the input to the memoryless erasure channel as the i^{th} row X_i^n of $\mathbf{D}^{(1)}$. The output $\bar{Y}^n$ is the matching row of $\bar{\mathbf{D}}^{(2)}$. For our row matching algorithm, we match the l^{th} row $\bar{Y}_l^n$ of $\bar{\mathbf{D}}^{(2)}$ with the i^{th} row X_i^n of $\mathbf{D}^{(1)}$, if X_i^n is the only row of $\mathbf{D}^{(1)}$ jointly ε-typical [49, Chapter 3] with $\bar{Y}_l^n$ with respect to p_{X^n,Y^n}, where

$$p_{X^n,\bar{Y}^n}(x^n, \bar{y}^n) = p_{X^n}(x^n) \prod_{j=1}^n q_{Y|X}(\bar{y}_j|x_j) \tag{55}$$

where X^n denotes the Markov chain of length n with probability transition matrix $\mathbf{P}$. This results in $\hat{\Theta}_n(1) = l$. Otherwise, declare *collision error*.

Similar to (24), using the union bound and the generalized AEP, the total probability of error of this scheme can be bounded as follows

$$P_e \leq \mu_n + \varepsilon + 2^{n(R-\bar{I}(X;\bar{Y})+3\varepsilon)} \tag{56}$$

Since m_n is exponential in n, by Lemma 3, $\mu_n \to 0$ as $n \to \infty$. Thus

$$P_e < 3\varepsilon \text{ as } n \to \infty \tag{57}$$

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

$$R < \bar{I}(X;\bar{Y}) \tag{58}$$

is achievable, by taking ε small enough. From [53, Corollary II.2] we have

$$\bar{I}(X;\bar{Y}) = (1-\delta)^2 \sum_{r=0}^{\infty} \delta^r H(X_0|X_{-r-1}) \tag{59}$$

where $H(X_0|X_{-r-1})$ is the entropy rate associated with the probability transition matrix $\mathbf{P}^{r+1}$.

Now, we argue that (48) can be proven via induction on r by taking (2) as a base case and observing that $\mathbf{U}^2 = \mathbf{U}$. Finally, plugging π and $\mathbf{P}^{r+1}$ directly into [49, Theorem 4.2.4] yields (49), concluding the achievability part of the proof. □

Next, we move on to prove the converse part of Theorem 2.

Proof of Converse of Theorem 2. Since the converse part of Theorem 1 holds for any seed size Λ_n, in the noiseless setting, we trivially have

$$C(\infty,\alpha) \leq \lim_{n\to\infty} \frac{I(X^n;Y^{K_n},S^n)}{n}. \tag{60}$$

Next, note that there is a bijective mapping between (Y^{K_n},S^n) and $(\bar{Y}^n,S^n)$. Therefore, we have

$$\begin{aligned} I(X^n;Y^{K_n},S^n) &= I(X^n;\bar{Y}^n,S^n) && (61)\\ &= I(X^n;\bar{Y}^n) + I(X^n;S^n|\bar{Y}^n) && (62)\\ &= I(X^n;\bar{Y}^n) && (63) \end{aligned}$$

where (63) follows from the independence of S^n and X^n conditioned on $\bar{Y}^n$. This is because since $\bar{Y}^n$ is stripped of all extra replicas, from $(X^n,\bar{Y}^n)$ we can only infer the zeros of S^n, which is already known through $\bar{Y}^n$ via erasure symbols. Thus, we have

$$C(\infty,\alpha) \leq \bar{I}(X;\bar{Y}) \tag{64}$$

where $\bar{I}(X;\bar{Y})$ is defined in (59), concluding the proof of the converse part. □

IV. Matching Capacity For Independent Repetition

In this section, we investigate the upper and the lower bounds on the matching capacity $C(d_{\text{rep}},\alpha)$ for independent repetition ($W_n = 1$, $d_{\text{rep}} = 0$), where we assume a repetition pattern which is independent across all rows.

Due to the independence of the repetition pattern and the independence of the database rows, the seeds do not offer any additional information on the repetition pattern or row matching. Thus, we focus on the matching capacity $C(0,\alpha)$ in the regime with no seeds, *i.e.,* $\Lambda_n = 0$. Furthermore, for tractability, we focus on the special case where $\gamma = 0$, resulting in an *i.i.d.* database distribution $p_X(x) = u_x$, $\forall x \in \mathfrak{X}$.

We state our main result on the matching capacity for independent repetition in the following theorem:

Theorem 3. ***(Matching Capacity Bounds for Independent Repetition)*** *Consider a probability transition matrix* $\mathbf{P}$ *with* $\gamma = 0$*, a noise distribution* $p_{Y|X}$ *and a repetition distribution* p_S*. Then the matching capacity satisfies*

$$C(0,\alpha) \geq \left[\frac{\mathbb{E}[S]}{s_{\max}} H(X) - (1-\alpha\delta) H_b\left(\frac{\mathbb{E}[S]}{(1-\alpha\delta)s_{\max}}\right) - \mathbb{E}[S] H(X|Y)\right]^+ \tag{65}$$

$$C(0,\alpha) \leq \inf_{n\geq 1} \frac{1}{n} I(X^n; Y^{K_n}, A^n) \tag{66}$$

where δ *and* α *are the deletion and the deletion detection probabilities, respectively and* $s_{\max} \triangleq \max \operatorname{supp}(p_S)$*. Furthermore, for repetition distributions with* $\frac{1}{s_{\max}}\mathbb{E}[S] \geq \frac{1-\alpha\delta}{|\mathfrak{X}|}$*, the lower bound in equation* (65) *can be tightened as*

$$\begin{aligned} C(0,\alpha) \geq \Big[(1-\alpha\delta)H(X) - \Big(1-\alpha\delta-\frac{\mathbb{E}[S]}{s_{\max}}\Big)\min\{H(X), \log(|\mathfrak{X}|-1)\} \\ - (1-\alpha\delta) H_b\left(\frac{\mathbb{E}[S]}{(1-\alpha\delta)s_{\max}}\right) - \mathbb{E}[S] H(X|Y)\Big]^+ \end{aligned} \tag{67}$$

We note that the upper bound given in Theorem 3 (equation (66)) is an infimum over the column size n. Therefore, its evaluation for any $n \in \mathbb{N}$ yields an upper bound on the matching capacity.

With independent repetition, we cannot perform repetition detection as in Section III, and hence we are restricted to using a single-step rowwise matching scheme as done in [34]. This builds an analogy between database matching and channel decoding. In particular, our approach to database matching for independent repetition is related to decoding in the noisy synchronization channel [54].

We stress that there are several important differences between the database matching problem and the synchronization channel literature: *i)* In database matching the database distribution is fixed and cannot be designed or optimized, whereas in channel coding the main goal is to optimize the input distribution to find the channel capacity *ii)* the synchronization channel literature mostly focuses on code design with few works, such as [55], focusing on random codebook arguments for only a few types of synchronization errors such as deletion [55] and duplication [56] and finally *iii)* our database matching result provides an achievability argument for all repetition distributions with finite support, whereas the synchronization channel literature mainly focuses on some families of repetition distributions. As a result, for input-constrained noisy synchronization channels, our generalized random codebook argument, presented in Section IV-A, is novel and might be of independent interest.

In Section IV-A, we prove the achievability part of Theorem 3 (equation (65)) by proposing a rowwise matching scheme. Then, in Section IV-B we prove the converse part (equation (66)). Then, we present strictly tighter upper bounds for a special case with only deletions, *i.e.,* $s_{\max} = 1$.

A. Row Matching Scheme and Achievability

To prove the achievability, we consider the following matching scheme:

1) Given the i^{th} row $Y_i^{K_n}$ of $\mathbf{D}^{(2)}$ and the corresponding row A_i^n of the partial deletion location information $\mathbf{A}$, we discard the j^{th} column of $\mathbf{D}^{(1)}$ if $A_{i,j} = 1$, $\forall j \in [n]$ to obtain $\bar{\mathbf{D}}^{(1)}$ since it does not offer any additional information due to the independent nature of the database entries.

2) We convert the problem into a deletion-only one by elementwise repeating all the columns of $\bar{\mathbf{D}}^{(1)}$ $s_{\max}$ times, which we call "stretching by $s_{\max}$", to obtain $\tilde{\mathbf{D}}^{(1)}$. At this step, $Y_i^{K_n}$ can be seen as the output of the noisy deletion channel where the $\Theta_n^{-1}(i)^{\text{th}}$ row of $\tilde{\mathbf{D}}^{(1)}$ is the input.

3) We perform a generalized version of the decoding algorithm introduced in [36] for the noiseless deletions with deletion detection probability. Note that the latter itself is an extension of the one proposed in [55].

The full proof of the achievability part (equations (65) and (67)) via the matching scheme described above can be found in Appendix D.

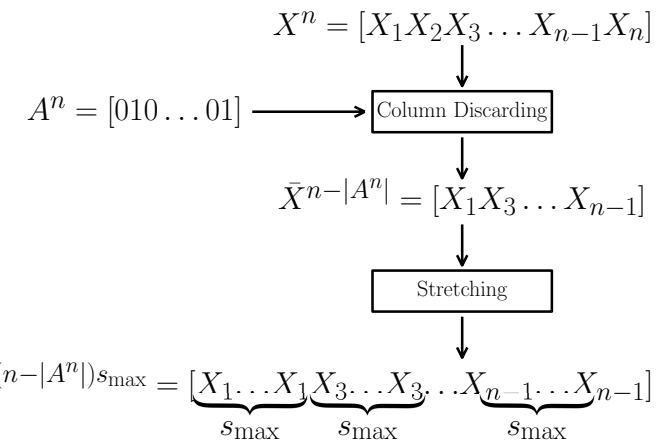

Fig. 4. An illustrative example of the column discarding and the stretching of X^n into $\tilde{X}^{(n-|A^n|)s_{\max}}$, for a given the deletion detection pattern A^n. First, we discard each know deleted element known from X^n to obtain $\bar{X}^{n-|A^n|}$. Then, each element of $\bar{X}^{(n-|A^n|)}$ is repeated $s_{\max}$ times to obtain $\tilde{X}^{(n-|A^n|)s_{\max}}$.

An illustrative example of the "stretching" is given in Figure 4. The idea behind this stretching is that since each entry can be repeated at most $s_{\max}$ times when we stretch X^n $s_{\max}$ times to obtain $\tilde{X}^{ns_{\max}}$, the output of the synchronization channel (before the noise $p_{Y|X}$) is guaranteed to be a subsequence of $\tilde{X}^{ns_{\max}}$. This way, we can convert the general noisy synchronization problem into a noisy deletion-only problem. We note that when $s_{\max}$ becomes large compared to the alphabet size $|\mathfrak{X}|$, the lower bound given in (65) goes to zero, even when $p_S(s_{\max})$ is very small.

For any repetition structure described in Definition 2, one can simply ignore the structure and apply the matching scheme described above. Therefore the achievable rate of Theorem 3 (equation (65)) is achievable for any repetition order.

Corollary 1. ***(Capacity Lower Bound for Arbitrary Repetition Order)*** *For any repetition order d_{rep} and any $\alpha \in [0,1]$, the matching capacity $C(d_{rep},\alpha)$ satisfies*

$$C(d_{rep},\alpha) \geq \left[\frac{\mathbb{E}[S]}{s_{\max}}H(X) - (1-\alpha\delta)H_b\left(\frac{\mathbb{E}[S]}{(1-\alpha\delta)s_{\max}}\right) - \mathbb{E}[S]H(X|Y)\right]^+ \tag{68}$$

B. Converse

In this subsection, we prove the converse part of Theorem 3 and evaluate the given upper bound for some special cases. First, we observe that by following the genie argument provided in the converse of Theorem 1, we can argue that Theorem 1 is an upper bound on $C(d_{\text{rep}},\alpha)$ for any α and for any d_{rep}.

Corollary 2. ***(Capacity Upper Bound for Arbitrary Repetition Order)*** *For any repetition order d_{rep} and any $\alpha \in [0,1]$, the matching capacity $C(d_{rep},\alpha)$ satisfies*

$$C(d_{rep},\alpha) \leq I(X;Y^S,S) \tag{69}$$

We next prove the converse of Theorem 3 (equation (66)). We then analytically evaluate this for some $n \in \mathbb{N}$ and we argue that the evaluated upper bounds are strictly tighter than that in Corollary 2.

Proof of Converse of Theorem 3. We start with the modified Fano's inequality used in Section III-D. Let

$$P_e \triangleq \Pr\left(\Theta_n(J) \neq \hat{\Theta}_n(J)\right), \quad J \sim \text{Unif}([m_n]) \tag{70}$$

Then, we have

$$H(\Theta_n) \leq 1 + m_n P_e \log m_n + I(\Theta_n;\mathbf{D}^{(1)},\mathbf{D}^{(2)},\mathbf{A}) \tag{71}$$

where

$$I(\Theta_n;\mathbf{D}^{(1)},\mathbf{D}^{(2)},\mathbf{A}) = I(\Theta_n;\mathbf{D}^{(1)}|\mathbf{D}^{(2)},\mathbf{A}) \tag{72}$$
$$\leq I(\Theta_n,\mathbf{D}^{(2)},\mathbf{A};\mathbf{D}^{(1)}) \tag{73}$$
$$= \sum_{i=1}^{m_n} I(X_i^n;Y_{\Theta_n(i)}^{K_n},A_{\Theta_n(i)}^n) \tag{74}$$
$$= m_n I(X^n;Y^{K_n},A^n) \tag{75}$$

where (74) and (75) follow from the fact that non-matching rows and their corresponding probabilistic side information on deletion locations are respectively independent and identically distributed. Following similar steps to Section III-D, we obtain

$$R \leq \lim_{n\to\infty} \frac{I(X^n;Y^{K_n},A^n)}{n} \tag{76}$$

whenever $P_e \to 0$ as $n \to \infty$.

Note that from Fekete's lemma [57], for any subadditive sequence $\{a_n\}_{n\in\mathbb{N}}$, we have

$$\lim_{n\to\infty} \frac{a_n}{n} = \inf_{n\geq 1} \frac{a_n}{n} \tag{77}$$

Therefore, it is sufficient to prove the subadditivity of $I(X^n;Y^{K_n},A^n)$.

Choose an arbitrary $r \in [n-1]$ and let $M_r \triangleq \sum_{j=1}^r S_j$ where S^n is the repetition pattern through which Y^{K_n} is obtained from X^n. Note that M_r denotes a marker, stating which part of Y^{K_n} depends on the first r elements of X^n, denoted by X_1^r. Therefore we have a bijective relation between (Y^{K_n},M_r) and $(Y_1^{\sum_{j=1}^r S_j},Y_{\sum_{j=1}^r S_j+1}^{K_n})$ where the subscripts and the superscripts denote the starting and the ending points of the vectors, respectively. Thus,

$$I(X^n;Y^{K_n},A^n) \leq I(X^n;Y^{K_n},M_r,A^n) \tag{78}$$
$$= I(X^n;Y_1^{\sum_{j=1}^r S_j},Y_{\sum_{j=1}^r S_j+1}^{K_n},A^n) \tag{79}$$
$$= I(X_1^r,X_{r+1}^n;Y_1^{\sum_{j=1}^r S_j},Y_{\sum_{j=1}^r S_j+1}^{K_n},A_1^r,A_{r+1}^n) \tag{80}$$
$$= I(X_1^r;Y_1^{\sum_{j=1}^r S_j},A_1^r) + I(X_{r+1}^n;Y_{\sum_{j=1}^r S_j+1}^{K_n},A_{r+1}^n) \tag{81}$$

where (81) follows from the fact that X^n and A^n have *i.i.d.* entries and the noise $p_{Y|X}$ acts independently on the entries. Thus, $I(X^n;Y^{K_n},A^n)$ is a subadditive sequence. Hence,

$$R \leq \inf_{n\geq 1} \frac{I(X^n;Y^{K_n},A^n)}{n} \tag{82}$$

whenever $P_e \to 0$ as $n \to \infty$, concluding the proof. □

We note that since the upper bound given in Theorem 3 is the infimum over all $n \geq 1$, its evaluation at any $n \in \mathbb{N}$ yields an upper bound on the matching capacity. In Corollaries 3 and 4, we analytically evaluate this upper bound at $n = 2$ under some assumptions on $p_{X,Y}$ when $s_{\max} = 1$, *i.e.,* when we only have deletions, and explicitly demonstrate the gap between the upper bounds given in Corollary 2 and Theorem 3.

First, we consider a noiseless deletion setting with arbitrary database distribution p_X in Corollary 3.

Corollary 3. ***(Upper Bound for Noiseless Deletion)*** *Consider a noiseless deletion setting where* $p_{Y|X}(y|x) = \mathbb{1}_{[x=y]}$, $\forall (x,y) \in \mathfrak{X}^2$ *and* $S \sim$ *Bernoulli*$(1-\delta)$. *Then for any input distribution* p_X, *we have*

$$C(0,\alpha) \leq \frac{1}{2} I(X^2;Y^K,A^2) \tag{83}$$
$$= (1-\delta)H(X) - (1-\alpha)\delta(1-\delta)(1-\hat{q}) \tag{84}$$

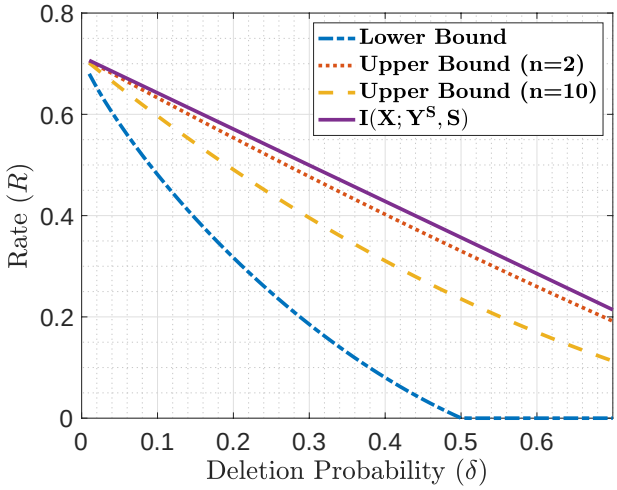

Fig. 5. The evaluation of the lower and upper bounds on the matching capacity for the binary noisy deletion case with $p_X \sim$Bernoulli($1/2$), $p_S \sim$ Bernoulli$(1-\delta)$, $\alpha = 0.7$ and $p_{Y|X} \sim$ BSC(0.05). The blue curve is the achievable rate stated in Theorem 3. The yellow and the red curves are the evaluations of the upper bound stated in Theorem 3, at $n = 10$ and $n = 2$, respectively. The purple curve shows the loose upper bound given in Corollary 2. We see that the gap between the lower and the upper bounds shrinks as n increases.

where $\hat{q} \triangleq \sum_{x\in\mathfrak{X}} p_X(x)^2$.

Proof. See Appendix E. □

Note that for any $\mathfrak{X}$ with $|\mathfrak{X}| \geq 2$ and $\alpha \in [0,1)$ the upper bound given in Corollary 3 is strictly lower than the one provided in Corollary 2 which is

$$I(X;Y,S) = (1-\delta)H(X). \tag{85}$$

Next, we consider a noisy deletion setting with binary X and arbitrary noise $p_{Y|X}$ in Corollary 4.

Corollary 4. ***(Upper Bound for Binary Noisy Deletion)*** *Consider a binary noisy deletion setting where* $X \sim$ *Bernoulli*(p) *and* $S \sim$ *Bernoulli*$(1-\delta)$. *Then, for any binary DMC* $p_{Y|X}$, *we have*

$$C(0,\alpha) \leq \frac{1}{2} I(X^2;Y^K,A^2) \tag{86}$$

$$= (1-\delta)I(X;Y) - 2(1-\alpha)\delta(1-\delta)p(1-p)I(U;V) \tag{87}$$

where U *and* V *are binary random variables with* $U \sim$ *Bernoulli*$(1/2)$ *and* $p_{V|U} = p_{Y|X}$.

Proof. See Appendix F. □

Again, for any $p \in (0,1)$ and $\alpha \in [0,1)$, the upper bound given in Corollary 4 is strictly lower than the one provided in Corollary 2 which is

$$I(X;Y,S) = (1-\delta)I(X;Y) \tag{88}$$

We note that the tighter upper bounds in Corollaries 3 and 4 become generalizations of the upper bound on the noiseless deletion channel mutual information, given in [58, Corollary 1]. Specifically, [58] considers noiseless deletion channel with *i.i.d.* Bernoulli inputs. Corollary 3 extends the results to noiseless deletion channels with arbitrary alphabet sizes. Furthermore, Corollary 4 extends the results to binary noisy deletion channels with arbitrary noise.

For the binary noisy case considered in Corollary 4, the numerical comparison of the lower bound and the two upper bounds on the matching capacity is provided in Figure 5. Note that the upper bound provided by Corollary 4 is not tight as it can be shown that a larger value of n gives a tighter upper bound, implying that the gap between the lower and the upper bounds in Theorem 3 is smaller than the one shown in Figure 5.

V. Extensions

In this section, we discuss extensions to the system model and results. Specifically, in Section V-A, we investigate the adversarial repetition case instead of random repetitions, where the repetitions are not due to random sampling of the time-indexed data, but due to a constrained privacy mechanism. In Section V-B, we consider the identical repetition model with no seeds. In Section V-C, we discuss the zero-rate regime, where the row size m_n is not necessarily exponential in the column size n, and derive conditions necessary for the detection algorithms discussed in Section III to work.

A. What If Repetitions Are Intentional?

So far, as stated in Definition 2, we have assumed that the repetitions occur randomly according to a discrete probability distribution p_S with finite integer support. In this subsection, we study the case of an adversary who controls the repetition pattern (under some constraints) to make matching as difficult as possible. This could arise for example where a privacy-preserving mechanism denies the sampling of the geolocation data when that data contains the most information about the users, such as their home addresses. We consider the adversarial setting under identical repetition assumption.

We stress that in the identical repetition setting, *i.e.,* $W_n = m_n$, the replicas either have no effect on the matching capacity as in the noiseless case (Theorem 2) or offer additional information acting as a repetition code of random length, in turn increasing the matching capacity (Theorem 1). Hence, it is expected that any adversary who tries to hinder the matching process to not allow the replication of entries. Therefore in the adversarial repetition setting, it is natural to focus on the deletion-only case. We assume an adversary with a δ-*deletion budget*, which can delete up to δ fraction of the columns, to maximize the mismatch probability. For tractability, we focus on the noiseless case with *i.i.d.* database entries. More formally, we assume $X_i \overset{\text{iid}}{\sim} p_X$ where

$$p_{Y|X}(y|x) = \mathbb{1}_{[y=x]}, \quad \forall (x,y) \in \mathfrak{X}^2 \tag{89}$$

Under these assumptions, we define the adversarial matching capacity as follows:

Definition 10. (Adversarial Matching Capacity) The *adversarial matching capacity* $C^{\text{adv}}(\delta)$ is the supremum of the set of all achievable rates corresponding to a database distribution p_X and an adversary with a δ-*deletion budget* when there is identical repetition. More formally,

$$C^{\text{adv}}(\delta) \triangleq \sup\{R : \forall I_{\text{del}} = (i_1, \ldots, i_{n\delta}) \subseteq [n], \Pr(\hat{\Theta}_n(J) \neq \Theta_n(J)) \overset{n\to\infty}{\longrightarrow} 0, J \sim \text{Uniform}([m_n]))\} \tag{90}$$

where the dependence of the matching scheme $\hat{\Theta}_n$ on the database growth rate R and the column deletion index set I_{del} is omitted for brevity.

Note that in this setting, although the deletions are not random, the matching error is still a random variable due to the random natures of $\mathbf{D}^{(1)}$ and Θ_n. In the proof of Theorem 4 below (Appendix G), we argue that in the adversarial setting, we can still convert deletions into erasures via the histogram-based repetition detection algorithm of Section III-E. After the detection part, we use the following matching scheme: We first remove deleted columns from $\mathbf{D}^{(1)}$, and then perform exact sequence matching.

We state our main result on the adversarial matching capacity in the following theorem:

Theorem 4. ***(Adversarial Matching Capacity)*** *Consider a database distribution p_X and an adversary with a* δ-deletion budget *when there is identical repetition ($W_n = m_n$). Then the adversarial matching capacity is*

$$C^{adv}(\delta) = \begin{cases} D(\delta \| 1-\hat{q}), & \text{if } \delta \leq 1-\hat{q} \\ 0, & \text{if } \delta > 1-\hat{q} \end{cases} \tag{91}$$

where $\hat{q} \triangleq \sum_{x\in\mathfrak{X}} p_X(x)^2$.

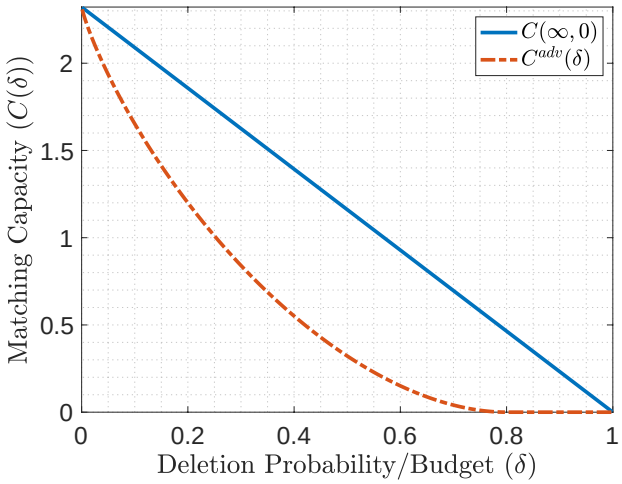

Fig. 6. Matching capacities C vs. deletion probability/budget (δ) when $X \sim \text{Unif}(\mathfrak{X})$, $\mathfrak{X} = [5]$. Notice that in this case $\hat{q} = 0.2$ and for $\delta > 1 - \hat{q} = 0.8$ the adversarial matching capacity $C^{\text{adv}}(\delta)$ is zero, while the matching capacity with random deletions $C(\infty,0)$ is positive.

Proof. See Appendix G. □

The matching capacities for random and adversarial deletions as a function of the deletion probability/budget are illustrated in Figure 6. Note that for $\delta > 1 - \hat{q}$, we have $C^{\text{adv}}(\delta) = 0$ whereas $C(\infty,0) = (1-\delta)H(X) > 0$. Furthermore, when $\delta \leq 1 - \hat{q}$ the matching capacity is significantly reduced when the column deletions are intentional rather than random.

B. What If There Were No Seeds?

In Section III, we assumed the availability of seeds with a seed size $\Lambda_n = \Omega(\log\log m_n)$. Now, we focus on the identical repetition scenario with no seeds.

Note that the replica detection algorithm of Section III-A does not require any seeds. Therefore in the seedless scenario, we can still detect the replicas with a vanishing probability of error. On the other hand, in the general noisy setting, the deletion detection algorithm of Section III-B necessitates seeds. Therefore, in the case of no seeds, we cannot perform deletion detection and we need to modify the matching scheme of Section III-C to obtain lower bounds on the matching capacity $C^{\text{seedless}}(\infty,0)$.

For tractability, we focus on the case with *i.i.d.* database entries, *i.e.,* $\gamma = 0$. More formally, we assume $X_i \overset{\text{iid}}{\sim} p_X$. Under this assumption, we state a lower bound on the unseeded matching capacity with identical repetition in the following theorem.

Theorem 5. ***(Seedless Matching Capacity with Identical Repetition)*** *Consider a database distribution p_X, a noise distribution $p_{Y|X}$, a repetition distribution p_S and an identical repetition pattern. Then, in the seedless case, the matching capacity $C^{seedless}(\infty,0)$ satisfies*

$$C^{seedless}(\infty,0) \geq \Big[I(X;Y^S,S) - H_b(\delta)\Big]^+ \tag{92}$$

$$C^{seedless}(\infty,0) \leq I(X;Y^S,S) \tag{93}$$

where $\delta \triangleq p_S(0)$ is the deletion probability, $S \sim p_S$ and Y^S has the following distribution conditioned on X such that

$$\Pr(Y^S = y^S | X = x) = \begin{cases} \prod_{j=1}^{S} p_{Y|X}(y_j|x) & \text{if } S > 0 \\ \mathbb{1}_{[y^S = E]} & \text{if } S = 0 \end{cases} \tag{94}$$

where E denotes the empty string.

Furthermore, for repetition distributions with $\delta \leq 1 - 1/|\mathfrak{X}|$, the lower bound can be tightened as

$$C^{no\ seed}(\infty,0) \geq \left[I(X;Y^S,S) + \delta[H(X) - \log(|\mathfrak{X}|-1)]^+ - H_b(\delta)\right]^+ \tag{95}$$

Proof. See Appendix H. □

We note that although the converse results of Theorems 1 and 5 match, the achievable rates differ by $H_b(\delta)$. In other words, Theorem 5 implies that the gap between the lower and the upper bounds on the seedless matching capacity is at most $H_b(\delta)$. We note that this gap is due to our inability to detect deletions in the achievability part. Hence, we conjecture that the lower bound in Theorem 5 is loose while the converse is tight. This is because in the noiseless setting, as discussed in Section III-E, deletion detection can be performed without seeds and the achievability bound is indeed improved and tight.

C. Zero-Rate Regime

In Section III, we considered at the matching capacity $C(\infty,0)$ for $\Lambda_n = \Omega(\log\log m_n)$ when the database growth rate R is positive. In other words, so far, we have assumed

$$\lim_{n\to\infty} \frac{1}{n} \log m_n > 0 \tag{96}$$

The detection algorithms we presented in Sections III-A through III-E depended on the row size m_n being large compared to the column size n. In this section, we further investigate these algorithms to derive the sufficient and/or necessary conditions on the relation between m_n and n in order for them to work in the zero-rate regime where

$$\lim_{n\to\infty} \frac{1}{n} \log m_n = 0. \tag{97}$$

Since $R = 0$, we define the non-asymptotic database growth rate R_n as

$$R_n \triangleq \frac{1}{n} \log m_n. \tag{98}$$

Here, $R = 0$ trivially implies $R_n \to 0$ as $n \to \infty$. Below we investigate the sufficient conditions on R_n such that the results of Sections III and IV hold.

1) Noisy Replica Detection: We consider the replica detection algorithm discussed in Section III-A. Note that the RHS of equation (142) of Appendix A has $2K - 2 \leq 2ns_{\max} = O(n)$ additive terms, each decaying exponentially in m_n. Thus, for a given average Hamming distance threshold $\tau \in (p_1, p_0)$ which is chosen based based on $\mathbf{P}$ and $p_{Y|X}$ and in turn constant with respect to n

$$m_n \geq \frac{\log(ns_{\max})}{\min\{D(\tau\|p_0), D(1-\tau\|1-p_1)\}} = \Theta(\log n) \tag{99}$$

is enough to ensure a vanishing replica detection error probability. In other words, as long as $m_n = \Omega(\log n)$ and in turn

$$R_n = \Omega\left(\frac{\log\log n}{n}\right) \tag{100}$$

our replica detection algorithm works.

2) Seeded Deletion Detection: We study the seeded deletion detection algorithm discussed in Section III-B. Note that we only run the deletion detection algorithm on the seeds $(\mathbf{G}^{(1)}, \mathbf{G}^{(2)})$ and not on the database pair $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)})$ directly, the relationship between m_n and n does not affect the success of the deletion detection. Thus, as long as the seed size $\Lambda_n = \Omega(\log n)$ our deletion detection algorithm works for any database growth rate, including the zero-rate regime. This in turn implies that $m_n \geq \Lambda_n = \Omega(\log n)$ and

$$R_n = \Omega\left(\frac{\log\log n}{n}\right). \tag{101}$$

3) Noiseless Joint Deletion-Replication Detection: We investigate the histogram-based joint deletion-replication detection algorithm introduced in Section III-E for the noiseless scenario. By Lemma 3, $m_n = \omega(n^4)$ is sufficient. Thus, as long as $\log m_n \geq 4\log n$, the histogram-based detection can be performed with a performance guarantee. In turn, for any

$$R_n = \Omega\left(\frac{\log n}{n}\right) \tag{102}$$

the histogram-based detection algorithm has a vanishing probability of error.

Therefore, in the noiseless setting, database growth rate $R_n = \Omega\left(\log n/n\right)$ provides enough granularity on the column histograms and we can perform detection with a decaying probability of error which then leads to asymptotically-zero mismatch probability.

Note that, for tractability, so far we have collapsed the databases into binary-valued ones. Further, in Lemma 3, we showed that for the collapsed databases $m_n = \omega(n^4)$ is enough for the asymptotic uniqueness of the column histograms. We now tighten this order relation for the special case where $\gamma = 0$ results in an *i.i.d.* database distribution $X_{i,j} \overset{\text{i.i.d.}}{\sim} p_X$ with support $\mathfrak{X}$.

Lemma 4. ***(Asymptotic Uniqueness of the Uncollapsed Histograms)*** *Consider an* i.i.d. *database distribution p_X. Let $H_j^{(1)}$ denote the histogram of the j^{th} column of $\mathbf{D}^{(1)}$. Then,*

$$\Pr\left(\exists i, j \in [n],\, i \neq j, H_i^{(1)} = H_j^{(1)}\right) \to 0 \text{ as } n \to \infty \tag{103}$$

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

Proof. See Appendix I. □

Note that in the binary setting the results of Lemmas 3 and 4 agree.

Lemma 4 implies that we only need a row size m_n polynomial in n to guarantee enough granularity for the uniqueness of $H_i^{(1)}$ and that the degree of the polynomial scales inversely with the alphabet size $|\mathfrak{X}|$. Furthermore, to demonstrate the tightness of this requirement of having $m_n = \omega(n^{\frac{4}{|\mathfrak{X}|-1}})$, we consider the special case where p_X is uniform over $\mathfrak{X}$. This leads to the following proposition:

Proposition 1. *Let $H_j^{(1)}$ denote the histogram of the j^{th} column of $\mathbf{D}^{(1)}$. If $p_X(x) = \frac{1}{|\mathfrak{X}|}$, $\forall x \in \mathfrak{X}$, then*

$$\Pr\left(\exists i, j \in [n],\, i \neq j, H_i^{(1)} = H_j^{(1)}\right) = n^2 m_n^{\frac{1-|\mathfrak{X}|}{2}} C_{|\mathfrak{X}|}(1 + o_n(1)) \tag{104}$$

where $C_{|\mathfrak{X}|} = (4\pi)^{\frac{1-|\mathfrak{X}|}{2}} |\mathfrak{X}|^{\frac{|\mathfrak{X}|}{2}}$.

Proof. See Appendix J. □

Proposition 1 states that in the setting with *i.i.d.* uniform database distribution, for the asymptotic uniqueness of the column histograms $m_n = \omega(n^{\frac{4}{|\mathfrak{X}|-1}})$ is not only sufficient but also necessary.

4) Independent Repetition Row Matching Scheme: In the independent repetition scenario, we have no detection algorithms which depend on the large-m_n assumption. Therefore, so long as the RHS of (65) is positive, any $R_n = o_n(1)$ is achievable. We stress that this observation trivially applies to the identical repetition case as well since one can simply ignore any underlying structure and perform the matching scheme given in Section IV-A.

VI. Conclusion

In this work, we have presented a unified information-theoretic foundation for database matching under noise and synchronization errors. We have showed that when the repetition pattern is constant across rows, the running Hamming distances between the consecutive columns of the correlated repeated database can be used to detect replicas. In addition, given seeds whose size grows double-logarithmic with the number of rows, a Hamming distance-based threshold testing, after an adequate remapping of database entries, 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. Furthermore, we have derived achievable database growth rates proposing a typicality-based matching scheme and a converse result for the setting where the repetition takes place entrywise, where we build analogy between database matching and synchronization channel decoding. We have also discussed some extensions, such as the adversarial column deletion setting rather then the random one.

Other natural extensions beyond those studied in this paper include the finite column size regime, where tools from finite-blocklength information theory could be useful, and practical algorithms with theoretical guarantees. An extensive analysis of the parallels between database matching under synchronization errors and two-dimensional synchronization channels [59], [60] and the construction of codes tailored to correct the error patterns investigated in this paper could be an interesting line of future work. Finally, one can extend our adversarial setting into a noisy one where the privacy-preserving mechanism not only deletes columns but also introduces intentional noise on the microdata, and investigate the adversarial matching capacity through a worst-case analysis.

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] G. Wondracek, T. Holz, E. Kirda, and C. Kruegel, "A practical attack to de-anonymize social network users," in *Proc. of IEEE Symposium on Security and Privacy*, 2010, pp. 223–238.

[9] J. Su, A. Shukla, S. Goel, and A. Narayanan, "De-anonymizing web browsing data with social networks," in *Proc. of the 26th International Conference on World Wide Web*, 2017, pp. 1261–1269.

[10] A. Shusterman, L. Kang, Y. Haskal, Y. Meltser, P. Mittal, Y. Oren, and Y. Yarom, "Robust website fingerprinting through the cache occupancy channel," in *28th USENIX Security Symposium (USENIX Security 19)*, 2019, pp. 639–656.

[11] B. Gulmezoglu, A. Zankl, T. Eisenbarth, and B. Sunar, "Perfweb: How to violate web privacy with hardware performance events," in *European Symposium on Research in Computer Security*. Springer, 2017, pp. 80–97.

[12] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda, "All your contacts are belong to us: Automated identity theft attacks on social networks," in *Proc. of the 18th International Conference on World Wide Web*, 2009, pp. 551–560.

[13] M. Srivatsa and M. Hicks, "Deanonymizing mobility traces: Using social network as a side-channel," in *Proc. of the 2012 ACM Conference on Computer and Communications Security*, 2012, pp. 628–637.

[14] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating Twitter users," in *Proc. of the 19th ACM International Conference on Information and Knowledge Management*, 2010, pp. 759–768.

[15] S. Kinsella, V. Murdock, and N. O'Hare, ""I'm eating a sandwich in Glasgow": Modeling locations with tweets," in *Proc. of the 3rd International Workshop on Search and Mining User-Generated Contents*, 2011, pp. 61–68.

[16] H. Kim, S. Lee, and J. Kim, "Inferring browser activity and status through remote monitoring of storage usage," in *Proc. of the 32nd Annual Conference on Computer Security Applications*, 2016, pp. 410–421.

[17] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific Reports*, vol. 3, no. 1, pp. 1–5, 2013.

[18] P. Erdos, A. Rényi *et al.*, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, no. 1, pp. 17–60, 1960.

[19] L. Babai, P. Erdos, and S. M. Selkow, "Random graph isomorphism," *SIAM Journal on computing*, vol. 9, no. 3, pp. 628–635, 1980.

[20] S. Janson, A. Rucinski, and T. Luczak, *Random Graphs*. John Wiley & Sons, 2011.

[21] T. Czajka and G. Pandurangan, "Improved random graph isomorphism," *Journal of Discrete Algorithms*, vol. 6, no. 1, pp. 85–92, 2008.

[22] L. Yartseva and M. Grossglauser, "On the performance of percolation graph matching," in *Proc. of the First ACM Conference on Online Social Networks*, 2013, pp. 119–130.

[23] P. Pedarsani, D. R. Figueiredo, and M. Grossglauser, "A Bayesian method for matching two similar graphs without seeds," in *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2013, pp. 1598–1607.

[24] M. Fiori, P. Sprechmann, J. Vogelstein, P. Musé, and G. Sapiro, "Robust multimodal graph matching: Sparse coding meets graph matching," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[25] V. Lyzinski, D. E. Fishkind, and C. E. Priebe, "Seeded graph matching for correlated Erdös-Rényi graphs," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3513–3540, 2014.

[26] E. Onaran, S. Garg, and E. Erkip, "Optimal de-anonymization in random graphs with community structure," in *2016 50th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2016, pp. 709–713.

[27] D. Cullina and N. Kiyavash, "Improved achievability and converse bounds for Erdos-Rényi graph matching," *ACM SIGMETRICS Performance Evaluation Review*, vol. 44, no. 1, pp. 63–72, 2016.

[28] A. Sanfeliu, R. Alquézar, J. Andrade, J. Climent, F. Serratosa, and J. Vergés, "Graph-based representations and techniques for image processing and image analysis," *Pattern Recognition*, vol. 35, no. 3, pp. 639–650, 2002.

[29] T. Galstyan, A. Minasyan, and A. Dalalyan, "Optimal detection of the feature matching map in presence of noise and outliers," *arXiv preprint arXiv:2106.07044*, 2021.

[30] B. Zhu, S. Chen, Y. Bai, H. Chen, N. Mukherjee, G. Vazquez, D. R. McIlwain, A. Tzankov, I. T. Lee, M. S. Matter *et al.*, "Robust single-cell matching and multi-modal analysis using shared and distinct features reveals orchestrated immune responses," *bioRxiv*, 2021.

[31] H. T. N. Tran, K. S. Ang, M. Chevrier, X. Zhang, N. Y. S. Lee, M. Goh, and J. Chen, "A benchmark of batch-effect correction methods for single-cell RNA sequencing data," *Genome Biology*, vol. 21, no. 1, pp. 1–32, 2020.

[32] J. Błażewicz, P. Formanowicz, M. Kasprzak, P. Schuurman, and G. J. Woeginger, "DNA Sequencing, Eulerian graphs, and the exact perfect matching problem," in *International Workshop on Graph-Theoretic Concepts in Computer Science*. Springer, 2002, pp. 13–24.

[33] 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.

[34] 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.

[35] 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.

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

[37] ——, "Seeded database matching under noisy column repetitions," in *2022 IEEE Information Theory Workshop (ITW)*, 2022.

[38] ——, "Matching of Markov databases under random column repetitions," in *2022 56th Asilomar Conference on Signals, Systems, and Computers*, 2022.

[39] 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.

[40] Z. K and B. Nazer, "Detecting correlated Gaussian databases," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 2064–2069.

[41] S. Chen, S. Jiang, Z. Ma, G. P. Nolan, and B. Zhu, "One-way matching of datasets with low rank signals," *arXiv preprint arXiv:2204.13858*, 2022.

[42] 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.

[43] 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.

[44] R. Bassily and A. Smith, "Causal erasure channels," in *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 2014, pp. 1844–1857.

[45] R. Gallager, "Sequential decoding for binary channels with noise and synchronization errors," *Lincoln Lab Group Report, October*, 1961.

[46] D. Fertonani, T. M. Duman, and M. F. Erden, "Bounds on the capacity of channels with insertions, deletions and substitutions," *IEEE Transactions on Communications*, vol. 59, no. 1, pp. 2–6, 2011.

[47] M. Rahmati and T. M. Duman, "Achievable rates for noisy channels with synchronization errors," *IEEE Transactions on Communications*, vol. 62, no. 11, pp. 3854–3863, 2014.

[48] A. Mandal, A. Chatterjee, and A. Thangaraj, "Noisy deletion, Markov codes and deep decoding," in *2020 National Conference on Communications (NCC)*, 2020, pp. 1–6.

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

[50] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. MIT press, 2022.

[51] R. Tamir, "Joint correlation detection and alignment of Gaussian databases," *arXiv preprint arXiv:2211.01069*, 2022.

[52] R. Gray and J. Kieffer, "Mutual information rate, distortion, and quantization in metric spaces," *IEEE Transactions on Information Theory*, vol. 26, no. 4, pp. 412–422, 1980.

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

[54] 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.
[55] 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.
[56] E. Drinea and M. Mitzenmacher, "Improved lower bounds for the capacity of i.i.d. deletion and duplication channels," *IEEE Transactions on Information Theory*, vol. 53, no. 8, pp. 2693–2714, 2007.
[57] M. Fekete, "Über die verteilung der wurzeln bei gewissen algebraischen gleichungen mit ganzzahligen koeffizienten," *Mathematische Zeitschrift*, vol. 17, no. 1, pp. 228–249, 1923.
[58] M. Drmota, W. Szpankowski, and K. Viswanathan, "Mutual information for a deletion channel," in *2012 IEEE International Symposium on Information Theory Proceedings*, 2012, pp. 2561–2565.
[59] L. Welter, R. Bitar, A. Wachter-Zeh, and E. Yaakobi, "Multiple criss-cross insertion and deletion correcting codes," *IEEE Transactions on Information Theory*, vol. 68, no. 6, pp. 3767–3779, 2022.
[60] E. Stylianou, L. Welter, R. Bitar, A. Wachter-Zeh, and E. Yaakobi, "Equivalence of insertion/deletion correcting codes for d-dimensional arrays," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 814–819.
[61] R. B. Ash, *Information Theory*. Courier Corporation, 2012.
[62] C. Burke and M. Rosenblatt, "A Markovian function of a Markov chain," *The Annals of Mathematical Statistics*, vol. 29, no. 4, pp. 1112–1122, 1958.
[63] V. Chvatal and D. Sankoff, "Longest common subsequences of two random sequences," *Journal of Applied Probability*, pp. 306–315, 1975.
[64] D. A. Brannan, *A First Course in Mathematical Analysis*. Cambridge University Press, 2006.
[65] L. B. Richmond and J. Shallit, "Counting Abelian squares," *arXiv preprint arXiv:0807.5028*, 2008.

APPENDIX A
PROOF OF LEMMA 1

Observe that since the rows of $\mathbf{D}^{(2)}$ are *i.i.d.* conditioned on the column repetition pattern S^n, the Hamming distance $d_H(C_j^{m_n}, C_{j+1}^{m_n})$ between consecutive columns $C_j^{m_n}$ and $C_{j+1}^{m_n}$ follows a Binomial distribution whose success parameter depends on whether $C_j^{m_n}$ and $C_{j+1}^{m_n}$ are noisy replicas or not.

Let H_1 denote the case where two random variables $Y_1, Y_2 \sim p_Y$ are noisy replicas and H_0 otherwise. Further, let

$$p_0 \triangleq \Pr(Y_1 \neq Y_2 | H_0) \tag{105}$$

$$p_1 \triangleq \Pr(Y_1 \neq Y_2 | H_1) \tag{106}$$

Then we have $d_H(C_j^{m_n}, C_{j+1}^{m_n}) \sim \text{Binom}(m_n, p_1)$ if $C_j^{m_n}$ and $C_{j+1}^{m_n}$ are noisy replicas and $d_H(C_j^{m_n}, C_{j+1}^{m_n}) \sim \text{Binom}(m_n, p_0)$ otherwise. Thus, proving that p_0 and p_1 are bounded away from one another will allow us to use the running Hamming distance based threshold test discussed in Section III-A.

Our goal is to prove that $p_0 > p_1$. First, we can formally rewrite p_0 as

$$p_0 = \sum_{x_1 \in \mathfrak{X}} \sum_{x_2 \in \mathfrak{X}} \sum_{y \in \mathfrak{X}} \Pr(X_1 = x_1) \Pr(X_2 = x_2 | X_1 = x_1) \Pr(Y_1 = y | X_1 = x_1) \Pr(Y_2 \neq y | X_2 = x_2) \tag{107}$$

$$= \sum_{x_1 \in \mathfrak{X}} \sum_{x_2 \in \mathfrak{X}} \sum_{y \in \mathfrak{X}} \Pr(X_1 = x_1) \Pr(X_2 = x_2 | X_1 = x_1) p_{Y|X}(y|x_1)[1 - p_{Y|X}(y|x_2)] \tag{108}$$

$$= \sum_{i=1}^{|\mathfrak{X}|} \sum_{j=1}^{|\mathfrak{X}|} \sum_{k=1}^{|\mathfrak{X}|} u_i P_{i,j} \, p_{Y|X}(k|i) \, [1 - p_{Y|X}(k|j)] \tag{109}$$

$$= \sum_{i=1}^{|\mathfrak{X}|} \sum_{j=1}^{|\mathfrak{X}|} \sum_{k=1}^{|\mathfrak{X}|} u_i \, [(1-\gamma)u_j + \gamma \delta_{ij}] \, p_{Y|X}(k|i) \, [1 - p_{Y|X}(k|j)] \tag{110}$$

$$= \sum_{i=1}^{|\mathfrak{X}|} \sum_{k=1}^{|\mathfrak{X}|} u_i \, [(1-\gamma)u_i + \gamma] \, p_{Y|X}(k|i) \, [1 - p_{Y|X}(k|i)] + \sum_{i=1}^{|\mathfrak{X}|} \sum_{j \neq i} \sum_{k=1}^{|\mathfrak{X}|} u_i \, [(1-\gamma)u_j] \, p_{Y|X}(k|i) \, [1 - p_{Y|X}(k|j)] \tag{111}$$

$$= (1-\gamma)\sum_{i=1}^{|\mathfrak{X}|}\sum_{j=1}^{|\mathfrak{X}|}\sum_{k=1}^{|\mathfrak{X}|} u_i\, u_j\, p_{Y|X}(k|i)\,[1-p_{Y|X}(k|j)] + \gamma\sum_{i=1}^{|\mathfrak{X}|}\sum_{k=1}^{|\mathfrak{X}|} u_i\, p_{Y|X}(k|i)\,[1-p_{Y|X}(k|i)] \tag{112}$$

$$= (1-\gamma)p_0' + \gamma p_1' \tag{113}$$

where

$$p_0' \triangleq \sum_{i=1}^{|\mathfrak{X}|}\sum_{j=1}^{|\mathfrak{X}|}\sum_{k=1}^{|\mathfrak{X}|} u_i\, u_j\, p_{Y|X}(k|i)\,[1-p_{Y|X}(k|j)] \tag{114}$$

$$p_1' \triangleq \sum_{i=1}^{|\mathfrak{X}|}\sum_{k=1}^{|\mathfrak{X}|} u_i\, p_{Y|X}(k|i)\,[1-p_{Y|X}(k|i)] \tag{115}$$

Similarly, we rewrite p_1 as

$$p_1 = \sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} \Pr(X=x)\Pr(Y_1=y|X=x)\Pr(Y_2\neq y|X=x) \tag{116}$$

$$= \sum_{i=1}^{|\mathfrak{X}|}\sum_{k=1}^{|\mathfrak{X}|} u_i\, p_{Y|X}(k|i)\,[1-p_{Y|X}(k|i)] \tag{117}$$

$$= p_1' \tag{118}$$

Thus, for any $\gamma\in[0,1)$ we have

$$p_0 > p_1 \iff p_0' > p_1' \tag{119}$$

Note that p_0' and p_1' would correspond to

$$p_0' = \Pr(Y_1\neq Y_2|H_0) \tag{120}$$

$$p_1' = \Pr(Y_1\neq Y_2|H_1) \tag{121}$$

if the entries $X_{i,j}$ of $\mathbf{D}^{(1)}$ were drawn *i.i.d.* from the stationary distribution π of $\mathbf{P}$, instead of a Markov process. Thus, to consider the *i.i.d.* database entries case, we introduce the discrete random variable W with

$$p_W(i) = u_i,\ \forall i\in\mathfrak{X} \tag{122}$$

$$p_{Y|W}(y|w) = p_{Y|X}(y|w),\ \forall (w,y)\in\mathfrak{X}^2 \tag{123}$$

Then, we can rewrite p_0' and p_1' as

$$p_0' = \sum_{w_1\in\mathfrak{X}}\sum_{w_2\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} p_W(w_1)p_W(w_2)p_{Y|W}(y|w_1)\left[1-p_{Y|W}(y|w_2)\right] \tag{124}$$

$$= \sum_{w_1\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} p_W(w_1)p_{Y|W}(y|w_1)\sum_{w_2\in\mathfrak{X}} p_W(w_2)\left[1-p_{Y|W}(y|w_2)\right] \tag{125}$$

$$= \sum_{w\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} p_W(w)p_{Y|W}(y|w)\left[1-p_Y(y)\right] \tag{126}$$

$$p_1' = \sum_{w\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} p_W(w)p_{Y|W}(y|w)\left[1-p_{Y|W}(y|w)\right] \tag{127}$$

Thus, we have

$$p_0' - p_1' = \sum_{w\in\mathfrak{X}}\sum_{y\in\mathfrak{X}} p_{W,Y}(w,y)\left[p_{Y|W}(y|w)-p_Y(y)\right]. \tag{128}$$

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

$$\psi(y) \triangleq \sum_{w\in\mathfrak{X}} p_W(w)\left[p_{Y|W}(y|w) - p_Y(y)\right]^2 \tag{129}$$

$$= \sum_{w\in\mathfrak{X}} p_W(w)\left[p_{Y|W}(y|w) - \sum_{z\in\mathfrak{X}} p_{Y|W}(y|z)p_W(z)\right]^2 \tag{130}$$

$$\geq 0 \tag{131}$$

where (131) follows from the non-negativity of the square term in the summation. It must be noted that $\psi(y) = 0$ only if $p_{Y|W}(y|w) = p_Y(y), \forall w \in \mathfrak{X}$ with $p_W(w) = u_w > 0$.

Expanding the square term, we obtain

$$\psi(y) = \sum_{w\in\mathfrak{X}} p_W(w)p_{Y|W}(y|w)^2 - 2p_Y(y)\sum_{w\in\mathfrak{X}} p_W(w)p_{Y|W}(y|w) + \sum_{w\in\mathfrak{X}} p_W(w)p_Y(y)^2 \tag{132}$$

$$= \sum_{w\in\mathfrak{X}} p_W(w)p_{Y|W}(y|w)^2 - 2p_Y(y)^2 + p_Y(y)^2 \tag{133}$$

$$= \sum_{w\in\mathfrak{X}} p_W(w)p_{Y|W}(y|w)^2 - p_Y(y)^2 \tag{134}$$

Next, we rewrite $p'_0 - p'_1$ as

$$p'_0 - p'_1 = \sum_{y\in\mathfrak{X}}\sum_{w\in\mathfrak{X}} p_{W,Y}(w,y)\left[p_{Y|W}(y|w) - p_Y(y)\right] \tag{135}$$

$$= \sum_{y\in\mathfrak{X}}\left[\left(\sum_{w\in\mathfrak{X}} p_W(w)p_{Y|W}(y|w)^2\right) - p_Y(y)^2\right] \tag{136}$$

$$= \sum_{y\in\mathfrak{X}} \psi(y) \tag{137}$$

$$\geq 0 \tag{138}$$

with $p'_0 - p'_1 = 0$ only when $p_{Y|W}(y|w) = p_Y(y), \forall (w,y) \in \mathfrak{X}^2$. In other words $p'_0 > p'_1$ and in turn $p_0 > p_1$ as long as the two databases are not independent.

We next 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 noisy replicas and B_j denote the event that the algorithm detects $C_j^{m_n}$ and $C_{j+1}^{m_n}$ as replicas. Via the union bound, we can upper bound the total probability of replica detection error as

$$\Pr(\bigcup_{j=1}^{K_n-1} E_j) \leq \sum_{j=1}^{K_n-1} \Pr(A_j^c)\Pr(B_j|A_j^c) + \Pr(A_j)\Pr(B_j^c|A_j) \tag{139}$$

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 the Chernoff bound [61, Lemma 4.7.2], we get

$$\Pr(B_j|A_j^c) \leq 2^{-m_n D(\tau\|p_0)} \tag{140}$$

$$\Pr(B_j^c|A_j) \leq 2^{-m_n D(1-\tau\|1-p_1)} \tag{141}$$

Thus, we get

$$\Pr(\bigcup_{j=1}^{K_n-1} E_j) \leq (K_n - 1)\left[2^{-m_n D(\tau\|p_0)} + 2^{-m_n D(1-\tau\|1-p_1)}\right] \tag{142}$$

Observe that since the RHS of (142) has $2K_n - 2 = O(n)$ terms decaying exponentially in m_n, for any $m_n = \omega(\log n)$ we have

$$\Pr(\bigcup_{j=1}^{K_n-1} E_j) \to 0 \text{ as } n \to \infty. \tag{143}$$

Finally observing that $n \sim \log m_n$ concludes the proof. □

APPENDIX B
PROOF OF LEMMA 2

Let $(\tilde{X}_{i,j}, \tilde{Y}_{i,j})$ be a pair of matching entries. Since the database distribution is stationary, WLOG, we can assume $(i,j) = (1,1)$. Now, given $(\tilde{X}_{1,1}, \tilde{Y}_{1,1})$, and the non-matching pair $(\tilde{X}_{1,j}, \tilde{Y}_{1,1})$ with $j-1 = r \neq 0$, we first prove the existence of such a bijective mapping σ such that for any $r \in [n-1]$

$$\Pr(\sigma(\tilde{Y}_{1,1}) \neq \tilde{X}_{1,1}) < \Pr(\sigma(\tilde{Y}_{1,1}) \neq \tilde{X}_{1,r+1}). \tag{144}$$

For given σ and $r \in [n-1]$ let

$$q_{0,\sigma}^{(r)} \triangleq \Pr(\sigma(\tilde{Y}_{1,1}) \neq \tilde{X}_{1,r+1}) \tag{145}$$

$$q_{1,\sigma} \triangleq \Pr(\sigma(\tilde{Y}_{1,1}) \neq \tilde{X}_{1,1}) \tag{146}$$

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

$$q_{0,\sigma}^{(r)} > q_{1,\sigma}, \forall r \in [n-1]. \tag{147}$$

We can rewrite $q_{0,\sigma}^{(r)}$ as

$$q_{0,\sigma}^{(r)} = \sum_{x_1 \in \mathfrak{X}} \sum_{x_2 \in \mathfrak{X}} \Pr(\tilde{X}_{1,1} = x_1) \Pr(\tilde{X}_{1,r+1} = x_2 | \tilde{X}_{1,1} = x_1) \Pr(\sigma(\tilde{Y}_{1,1}) \neq x_2 | \tilde{X}_{1,1} = x_1) \tag{148}$$

$$= \sum_{i=1}^{|\mathfrak{X}|} \sum_{j=1}^{|\mathfrak{X}|} u_i \, (\mathbf{P}^r)_{i,j} \, [1 - p_{Y|X}(\sigma^{-1}(j)|i)] \tag{149}$$

$$= \sum_{i=1}^{|\mathfrak{X}|} \sum_{j=1}^{|\mathfrak{X}|} u_i \, [(1-\gamma^r)u_j + \gamma^r \delta_{ij}] \, [1 - p_{Y|X}(\sigma^{-1}(j)|i)] \tag{150}$$

$$= (1-\gamma^r) \sum_{i=1}^{|\mathfrak{X}|} \sum_{j=1}^{|\mathfrak{X}|} u_i u_j \, [1 - p_{Y|X}(\sigma^{-1}(j)|i)] + \gamma^r \sum_{i=1}^{|\mathfrak{X}|} u_i \, [1 - p_{Y|X}(\sigma^{-1}(i)|i)] \tag{151}$$

$$= (1-\gamma^r) q'_{0,\sigma} + \gamma^r q'_{1,\sigma} \tag{152}$$

where

$$q'_{0,\sigma} \triangleq \sum_{i=1}^{|\mathfrak{X}|} \sum_{j=1}^{|\mathfrak{X}|} u_i u_j \, [1 - p_{Y|X}(\sigma^{-1}(j)|i)] \tag{153}$$

$$q'_{1,\sigma} \triangleq \sum_{i=1}^{|\mathfrak{X}|} u_i \, [1 - p_{Y|X}(\sigma^{-1}(i)|i)] \tag{154}$$

Similarly, we rewrite $q_{1,\sigma}$ as

$$q_{1,\sigma} = \sum_{x \in \mathfrak{X}} \Pr(\tilde{X}_{1,1} = x) \Pr(\sigma(\tilde{Y}_{1,1}) \neq x | \tilde{X}_{1,1} = x) \tag{155}$$

$$= \sum_{i=1}^{|\mathfrak{X}|} u_i [1 - p_{Y|X}(\sigma^{-1}(i)|i)] \tag{156}$$

$$= q'_{1,\sigma} \tag{157}$$

Thus, for any $\gamma \in [0,1)$, we have

$$\exists \sigma, \forall r \in [n-1],\, q_{0,\sigma}^{(r)} > q_{1,\sigma} \iff \exists \sigma,\, q'_{0,\sigma} > q'_{1,\sigma} \tag{158}$$

Note that $q'_{0,\sigma}$ and $q'_{1,\sigma}$ correspond to

$$q'_{0,\sigma} = \Pr(\sigma(\tilde{Y}_{1,1}) \neq \tilde{X}_{1,j}), \quad j \neq 1 \tag{159}$$

$$q'_{1,\sigma} = \Pr(\sigma(\tilde{Y}_{1,1}) \neq \tilde{X}_{1,1}) \tag{160}$$

if the entries $\tilde{X}_{i,j}$ of $\mathbf{G}^{(1)}$ were drawn *i.i.d.* from the distribution $\pi = [u_1, \ldots, u_{|\mathfrak{X}|}]$, instead of a Markov process. Thus, we recall the discrete random variable W, defined in equations (122)-(123), with

$$p_W(i) = u_i, \forall i \in \mathfrak{X} \tag{161}$$

$$p_{Y|W}(y|w) = p_{Y|X}(y|w), \forall (w,y) \in \mathfrak{X}^2 \tag{162}$$

Then, we can rewrite $q'_{0,\sigma}$ and $q'_{1,\sigma}$ as

$$q'_{0,\sigma} = \sum_{w_1 \in \mathfrak{X}} \sum_{w_2 \in \mathfrak{X}} p_W(w_1) p_W(w_2) [1 - p_{Y|X}(\sigma^{-1}(w_2)|w_1)] \tag{163}$$

$$q'_{1,\sigma} = \sum_{w \in \mathfrak{X}} p_W(w) [1 - p_{Y|W}(\sigma^{-1}(w)|w)] \tag{164}$$

We first prove the following:

$$\sum_{\sigma} q'_{0,\sigma} - q'_{1,\sigma} = 0 \tag{165}$$

where the summation is over all permutations of $\mathfrak{X}$. For brevity, let

$$Q_{i,j} \triangleq p_{Y|W}(j|i) \quad \forall i, j \in \mathfrak{X} \tag{166}$$

Note that from (166), we have

$$\sum_{j=1}^{|\mathfrak{X}|} Q_{i,j} = 1 \quad \forall i \in \mathfrak{X} \tag{167}$$

$$\sum_{i=1}^{|\mathfrak{X}|} \sum_{j=1}^{|\mathfrak{X}|} Q_{i,j} = |\mathfrak{X}| \tag{168}$$

Taking the sum over all σ, we obtain

$$\sum_{\sigma} q'_{0,\sigma} - q'_{1,\sigma} = \sum_{\sigma} \sum_{i=1}^{|\mathfrak{X}|} \sum_{j=1}^{|\mathfrak{X}|} p_W(i) p_W(j) Q_{i,\sigma^{-1}(j)} - \sum_{\sigma} \sum_{i=1}^{|\mathfrak{X}|} p_W(i) Q_{i,\sigma^{-1}(i)} \tag{169}$$

Combining (167)-(169), we now show that both terms on the RHS of (169) are equal to $(|\mathfrak{X}|-1)!$. We first look at the second term on the RHS of (169).

$$\sum_{\sigma} \sum_{i=1}^{|\mathfrak{X}|} p_W(i) Q_{i,\sigma^{-1}(i)} = \sum_{i=1}^{|\mathfrak{X}|} p_W(i) \sum_{\sigma} Q_{i,\sigma^{-1}(i)} \tag{170}$$

$$= (|\mathfrak{X}|-1)! \sum_{j=1}^{|\mathfrak{X}|} \sum_{i=1}^{|\mathfrak{X}|} p_W(i) Q_{i,j} \tag{171}$$

$$= (|\mathfrak{X}|-1)! \sum_{i=1}^{|\mathfrak{X}|} \sum_{j=1}^{|\mathfrak{X}|} p_{W,Y}(i,j) \tag{172}$$

$$= (|\mathfrak{X}|-1)! \tag{173}$$

where (171) follows from the fact that for any $j \in \mathfrak{X}$, we have exactly $(|\mathfrak{X}|-1)!$ permutations assigning j to i (or equivalently $\sigma^{-1}(i) = j$).

Now we look at the first term.

$$\sum_{\sigma}\sum_{i=1}^{|\mathfrak{X}|}\sum_{j=1}^{|\mathfrak{X}|} p_W(i)p_W(j)Q_{i,\sigma^{-1}(j)} = \sum_{i=1}^{|\mathfrak{X}|}\sum_{j=1}^{|\mathfrak{X}|} p_W(i)p_W(j)\sum_{\sigma} Q_{i,\sigma^{-1}(j)} \tag{174}$$

$$= \sum_{i=1}^{|\mathfrak{X}|}\sum_{j=1}^{|\mathfrak{X}|} p_W(i)p_W(j)(|\mathfrak{X}|-1)!\sum_{k=1}^{|\mathfrak{X}|} Q_{i,k} \tag{175}$$

$$= (|\mathfrak{X}|-1)! \tag{176}$$

Again, (175) follows from the fact that for each $k \in \mathfrak{X}$, there are exactly $(|\mathfrak{X}|-1)!$ permutations σ which map k to j (or equivalently $\sigma^{-1}(j) = k$).

Thus, we have showed that both terms on the RHS of (169) are equal to $(|\mathfrak{X}|-1)!$, proving (165). Now, we only need to show that

$$\exists \sigma \quad q'_{0,\sigma} - q'_{1,\sigma} \neq 0. \tag{177}$$

This is because unless $q'_{0,\sigma} - q'_{1,\sigma} = 0 \ \forall \sigma$, due to (165), we automatically have a σ such that this difference is strictly positive. This follows from the fact if $\exists \sigma \ q'_{0,\sigma} - q'_{1,\sigma} \neq 0$, we have either

- $q'_{0,\sigma} - q'_{1,\sigma} > 0$, which is the desired result, or
- $q'_{0,\sigma} - q'_{1,\sigma} < 0$, which from (169) requires the existence of another permutation $\tilde{\sigma}$ with $q'_{0,\tilde{\sigma}} - q'_{1,\tilde{\sigma}} > 0$.

We will prove (177) by arguing that

$$q'_{0,\sigma} - q'_{1,\sigma} = 0 \quad \forall \sigma \iff p_{Y|W}(y|w) = p_Y(y) \quad \forall (w,y) \in \mathfrak{X}^2 \tag{178}$$

which contradicts our $p_{Y|X} \neq p_Y$ assumption.

We first prove the "only if" part. Suppose $p_{Y|W}(y|w) = p_Y(y)$, $\forall (w,y) \in \mathfrak{X}^2$. In other words, $Q_{i,k} = Q_{j,k}$, $\forall (i,j,k) \in \mathfrak{X}^3$. Then for any σ, we have

$$q'_{0,\sigma} = \sum_{i=1}^{|\mathfrak{X}|}\sum_{j=1}^{|\mathfrak{X}|} p_W(i)p_W(j)Q_{i,\sigma^{-1}(j)} \tag{179}$$

$$= \sum_{i=1}^{|\mathfrak{X}|}\sum_{j=1}^{|\mathfrak{X}|} p_W(i)p_W(j)Q_{k,\sigma^{-1}(j)}, \quad k \neq i \tag{180}$$

$$= \sum_{j=1}^{|\mathfrak{X}|} p_W(j)Q_{k,\sigma^{-1}(j)} \tag{181}$$

$$= \sum_{j=1}^{|\mathfrak{X}|} p_W(j)Q_{j,\sigma^{-1}(j)} \tag{182}$$

$$= q'_{1,\sigma} \tag{183}$$

finishing the proof of the "only if" part.

Now, we prove the "if" part. Suppose the LHS of (178) holds. In other words, for any σ

$$\sum_{i=1}^{|\mathfrak{X}|} p_W(i)Q_{i,\sigma^{-1}(i)} = \sum_{i=1}^{|\mathfrak{X}|}\sum_{j=1}^{|\mathfrak{X}|} p_W(i)p_W(j)Q_{i,\sigma^{-1}(j)} \tag{184}$$

First, we look at the binary case $\mathfrak{X}=\{1,2\}$. In this case, we obtain

$$\begin{aligned} p_W(1)Q_{1,1}+p_W(2)Q_{2,2} &= p_W(1)^2Q_{1,1}+p_W(1)p_W(2)Q_{1,2} \\ &\quad +p_W(2)p_W(1)Q_{2,1}+p_W(2)^2Q_{2,2} && (185)\\ Q_{1,1}+Q_{2,2} &= Q_{1,2}+Q_{2,1} && (186)\\ Q_{1,1}+Q_{2,2} &= 1-Q_{1,1}+1-Q_{2,2} && (187)\\ Q_{1,1}+Q_{2,2} &= 1 && (188)\end{aligned}$$

for the identity permutation. This implies that $Q_{1,1}=Q_{2,1}$ and $Q_{1,2}=Q_{2,2}$ and this in turn implies $p_{Y|W}(y|w)=p_Y(y)\ \forall(w,y)\in\mathfrak{X}^2$, concluding the proof for the binary case.

Now, we investigate the larger alphabet sizes ($|\mathfrak{X}|\geq 3$). Since the equality holds for all σ, we now carefully select some one-cycle permutations σ to construct a system of linear equations.

Let σ_{id} be the identity permutation and $\sigma_{i-j},\sigma_{i-k},\sigma_{i-j-k}$ denote the one-cycle permutations with the respective cycles $(i\,j)$, $(i\,k)$ and $(i\,j\,k)$ for some distinct (i,j,k) triplet. For the rest of this proof, we will jointly solve the system of equations put forward by these permutations.

Recall that $p_W(l)=u_l$, $\forall l\in\mathfrak{X}$. Then, σ_{id} leads to

$$\begin{aligned} u_iQ_{i,i}+u_jQ_{j,j}+u_kQ_{k,k}+\sum_{l\neq i,j,k}u_lQ_{l,l} & \\ = u_i\sum_{t=1}^{|\mathfrak{X}|}u_tQ_{t,i}+u_j\sum_{t=1}^{|\mathfrak{X}|}u_tQ_{t,j}+u_k\sum_{t=1}^{|\mathfrak{X}|}u_tQ_{t,k}+\sum_{l\neq i,j,k}u_l\sum_{t=1}^{|\mathfrak{X}|}u_tQ_{t,l} && (189)\end{aligned}$$

Similarly, σ_{i-j} leads to

$$\begin{aligned} u_iQ_{i,j}+u_jQ_{j,i}+u_kQ_{k,k}+\sum_{l\neq i,j,k}u_lQ_{l,l} & \\ = u_i\sum_{t=1}^{|\mathfrak{X}|}u_tQ_{t,j}+u_j\sum_{t=1}^{|\mathfrak{X}|}u_tQ_{t,i}+u_k\sum_{t=1}^{|\mathfrak{X}|}u_tQ_{t,k}+\sum_{l\neq i,j,k}u_l\sum_{t=1}^{|\mathfrak{X}|}u_tQ_{t,l} && (190)\end{aligned}$$

When we subtract (190) from (189), we obtain

$$u_i(Q_{i,i}-Q_{i,j})-u_j(Q_{j,i}-Q_{j,j})=(u_i-u_j)\sum_{t=1}^{|\mathfrak{X}|}u_t(Q_{t,i}-Q_{t,j}) \tag{191}$$

Equivalently, we have

$$\begin{aligned} p_{W,Y}(i,i)-p_{W,Y}(i,j)-p_{W,Y}(j,i)+p_{W,Y}(j,j) & \\ = p_W(i)p_Y(i)-p_W(i)p_Y(j)-p_W(j)p_Y(i)+p_W(j)p_Y(j) && (192)\end{aligned}$$

Following the same steps, from σ_{i-k} we get

$$\begin{aligned} p_{W,Y}(i,i)-p_{W,Y}(i,k)-p_{W,Y}(k,i)+p_{W,Y}(k,k) & \\ = p_W(i)p_Y(i)-p_W(i)p_Y(k)-p_W(k)p_Y(i)+p_W(k)p_Y(k) && (193)\end{aligned}$$

We can rearrange the terms in (193) to obtain

$$\begin{aligned} p_{W,Y}(i,k) &= p_{W,Y}(i,i)-p_{W,Y}(k,i)+p_{W,Y}(k,k) \\ &\quad -p_W(i)p_Y(i)+p_W(i)p_Y(k)+p_W(k)p_Y(i)-p_W(k)p_Y(k) && (194)\end{aligned}$$

Furthermore, σ_{i-j-k} gives us

$$u_iQ_{i,k}+u_jQ_{j,i}+u_kQ_{k,j}+\sum_{l\neq i,j,k}u_lQ_{l,l} \\ =u_i\sum_{t=1}^{|\mathfrak{X}|}u_tQ_{t,k}+u_j\sum_{t=1}^{|\mathfrak{X}|}u_tQ_{t,i}+u_k\sum_{t=1}^{|\mathfrak{X}|}u_tQ_{t,j}+\sum_{l\neq i,j,k}u_l\sum_{t=1}^{|\mathfrak{X}|}u_tQ_{t,l} \tag{195}$$

Subtracting (195) from (190) yields

$$u_i(Q_{i,j}-Q_{i,k})+u_k(Q_{k,k}-Q_{k,j})=(u_i-u_k)\left[\sum_{t=1}^{|\mathfrak{X}|}u_tQ_{t,j}-\sum_{t=1}^{|\mathfrak{X}|}u_tQ_{t,k}\right] \tag{196}$$

Equivalently,

$$p_{W,Y}(i,j)-p_{W,Y}(i,k)-p_{W,Y}(k,j)+p_{W,Y}(k,k) \\ =p_W(i)p_Y(j)-p_W(i)p_Y(k)-p_W(k)p_Y(j)+p_W(k)p_Y(k) \tag{197}$$

Plugging $p_{W,Y}(i,k)$ from (194) into (197) yields

$$p_{W,Y}(i,j)-p_{W,Y}(i,i)-p_{W,Y}(k,j)+p_{W,Y}(k,i) \\ =p_W(i)p_Y(j)-p_W(i)p_Y(i)-p_W(k)p_Y(j)+p_W(k)p_Y(i) \tag{198}$$

Taking a summation over k in (198) gives us

$$|\mathfrak{X}|p_{W,Y}(i,j)-|\mathfrak{X}|p_{W,Y}(i,i)-p_Y(j)+p_Y(i) \\ =|\mathfrak{X}|p_W(i)p_Y(j)-|\mathfrak{X}|p_W(i)p_Y(i)-p_Y(j)+p_Y(i) \tag{199}$$

$$p_{W,Y}(i,j)-p_{W,Y}(i,i)=p_W(i)p_Y(j)-p_W(i)p_Y(i) \tag{200}$$

Similarly, taking a summation over j in (200) yields

$$p_W(i)-|\mathfrak{X}|p_{W,Y}(i,i)=p_W(i)-|\mathfrak{X}|p_W(i)p_Y(i) \tag{201}$$

$$p_{W,Y}(i,i)=p_W(i)p_Y(i) \tag{202}$$

Plugging (202) into (200) yields

$$p_{W,Y}(i,j)-p_{W,Y}(i,i)=p_W(i)p_Y(j)-p_W(i)p_Y(i) \tag{203}$$

$$p_{W,Y}(i,j)=p_W(i)p_Y(j) \tag{204}$$

Note that i and j are chosen arbitrarily. Therefore the condition given in (184) implies that $p_{Y|W}(y|w)=p_Y(y)$, $\forall(w,y)\in\mathfrak{X}^2$, concluding the proof of the "if" part.

Hence, we have proved (177). Thus, there exists a deterministic bijective mapping σ satisfying $q'_{0,\sigma}>q'_{1,\sigma}$ and in turn $q^{(r)}_{0,\sigma}>q'_{1,\sigma}$, $\forall r\in[n-1]$.

Now choose such a mapping σ and note that for any $\gamma\in[0,1)$

$$q^{(r)}_{0,\sigma}-q'_{1,\sigma}=(1-\gamma^r)[q'_0(\sigma)-q'_1(\sigma)] \tag{205}$$

$$\geq(1-\gamma)[q'_0(\sigma)-q'_1(\sigma)],\ \forall r\in[n-1] \tag{206}$$

$$>0,\ \forall r\in[n-1] \tag{207}$$

Next, define

$$q^{\min}_{0,\sigma}\triangleq(1-\gamma)q'_{0,\sigma}+\gamma q'_{1,\sigma} \tag{208}$$

and choose a $\bar{\tau}\in\left(q'_{1,\sigma},q^{\min}_{0,\sigma}\right)$ bounded away from both ends of the interval.

Let $\hat{K}_n \triangleq n - \sum_{j=1}^{n} I_j$ and L_j denote the j^{th} 0 in I^n, $j = 1, \ldots, \hat{K}_n$. In other words, L_j holds the index of the j^{th} retained column $C_j^{(2)}(\sigma)$ of $\tilde{\mathbf{G}}_\sigma^{(2)}$ in $\mathbf{G}^{(1)}$. Similarly, for i with $I_i = 0$, let $R_i \triangleq i - \sum_{l=1}^{i} I_l$ store the index of $C_i^{(1)}$ in $\tilde{\mathbf{G}}_\sigma^{(2)}$.

Now note that when we have $I_i = 1$, $d_H(C_i^{(1)}, C_j^{(2)}(\sigma)) \sim \text{Binom}(\Lambda_n, q_{0,\sigma}^{(|i-L_j|)})$ and when $I_i = 0$, $d_H(C_i^{(1)}, C_{R_i}^{(2)}(\sigma)) \sim \text{Binom}(\Lambda_n, q'_{1,\sigma})$.

Next, we write the misdetection probability $P_{e,i}$ of $C_i^{(1)}$ as

$$\begin{aligned} P_{e,i} &= \Pr\left(\exists j \in [\hat{K}_n] : \Delta_{i,j}(\sigma) \le \Lambda_n \bar{\tau}, I_i = 1\right) \\ &\quad + \Pr\left(\forall j \in [\hat{K}_n] : \Delta_{i,j}(\sigma) > \Lambda_n \bar{\tau}, I_i = 0\right) \qquad (209) \\ &\le \Pr\left(\exists j \in [\hat{K}_n] : \Delta_{i,j}(\sigma) \le \Lambda_n \bar{\tau}, I_i = 1\right) \\ &\quad + \Pr\left(\Delta_{i,R_i}(\sigma) > \Lambda_n \bar{\tau}, I_i = 0\right) \qquad (210) \end{aligned}$$

where

$$\Delta_{i,j}(\sigma) \triangleq d_H(C_i^{(1)}, C_j^{(2)}(\sigma)). \qquad (211)$$

From the union bound and Chernoff bound [61, Lemma 4.7.2], we obtain

$$\begin{aligned} P_{e,i} &\le \sum_{j=1}^{\hat{K}_n} \Pr\left(\Delta_{i,j}(\sigma) \le \Lambda_n \bar{\tau}, I_i = 1\right) + \Pr\left(\Delta_{i,R_i}(\sigma) > \Lambda_n \bar{\tau}, I_i = 0\right) \qquad (212) \\ &\le \sum_{j=1}^{\hat{K}_n} 2^{-\Lambda_n D(\bar{\tau} \| q_{0,\sigma}^{(|i-L_j|)})} + 2^{-\Lambda_n D(1-\bar{\tau} \| 1-q'_{1,\sigma})} \qquad (213) \end{aligned}$$

It is straightforward to show that $D(\bar{\tau}\|p)$ is an increasing function of p for $p > \bar{\tau}$. Thus $\forall i \in [n], j \in [\hat{K}_n]$, we have

$$\begin{aligned} q_{0,\sigma}^{(|i-L_j|)} &\ge q'_{0,\sigma} \qquad (214) \\ D(\bar{\tau} \| q_{0,\sigma}^{(|i-L_j|)}) &\ge D(\bar{\tau} \| q_{0,\sigma}^{\min}) \qquad (215) \\ 2^{-\Lambda_n D(\bar{\tau} \| q_{0,\sigma}^{(|i-L_j|)})} &\le 2^{-\Lambda_n D(\bar{\tau} \| q_{0,\sigma}^{\min})} \qquad (216) \end{aligned}$$

Thus, we have

$$\begin{aligned} P_{e,i} &\le \sum_{j=1}^{\hat{K}_n} 2^{-\Lambda_n D(\bar{\tau} \| q_{0,\sigma}^{(|i-L_j|)})} + 2^{-\Lambda_n D(1-\bar{\tau} \| 1-q'_{1,\sigma})} \qquad (217) \\ &\le \sum_{j=1}^{\hat{K}_n} 2^{-\Lambda_n D(\tau \| q_{0,\sigma}^{\min})} + 2^{-\Lambda_n D(1-\tau \| 1-q'_{1,\sigma})} \qquad (218) \\ &= \hat{K}_n 2^{-\Lambda_n D(\tau \| q_{0,\sigma}^{\min})} + 2^{-\Lambda_n D(1-\tau \| 1-q'_{1,\sigma})} \qquad (219) \end{aligned}$$

Thus, by simple union bound the total misdetection probability $P_{e,total}$ can be bounded as

$$\begin{aligned} P_{e,total} &\le \sum_{i=1}^{n} P_{e,i} \qquad (220) \\ &\le \sum_{i=1}^{n} \hat{K}_n 2^{-\Lambda_n D(\bar{\tau} \| q_{0,\sigma}^{\min})} + 2^{-\Lambda_n D(1-\bar{\tau} \| 1-q'_{1,\sigma})} \qquad (221) \\ &= n\hat{K}_n 2^{-\Lambda_n D(\bar{\tau} \| q_{0,\sigma}^{\min})} + n2^{-\Lambda_n D(1-\bar{\tau} \| 1-q'_{1,\sigma})} \qquad (222) \\ &\le n^2 2^{-\Lambda_n D(\bar{\tau} \| q_{0,\sigma}^{\min})} + n2^{-\Lambda_n D(1-\bar{\tau} \| 1-q'_{1,\sigma})} \qquad (223) \end{aligned}$$

Hence, $P_{e,total} \to 0$ as $n \to \infty$ if the seed size Λ_n satisfies

$$\Lambda_n D(\bar{\tau} \| q_{0,\sigma}^{\min}) - 2\log n > 0 \tag{224}$$

$$\Lambda_n D(1-\bar{\tau} \| 1-q'_{1,\sigma}) - \log n > 0 \tag{225}$$

Thus any seed size Λ_n satisfying

$$\Lambda_n > \frac{\log n}{\min\left\{\frac{1}{2}D(\bar{\tau} \| q_{0,\sigma}^{\min}), D(1-\bar{\tau} \| 1-q'_{1,\sigma})\right\}} \tag{226}$$

is sufficient to drive $P_{e,total}$ to 0. Thus a seed size $\Lambda_n = \Omega(\log n) = \Omega(\log\log m_n)$ is enough for successful deletion detection. □

APPENDIX C
PROOF OF LEMMA 3

First, observe that from [62, Theorem 3] and (2), the rows of the collapsed database $\tilde{\mathbf{D}}^{(1)}$ become *i.i.d.* first-order stationary binary Markov chains, with the following probability transition matrix and stationary distribution:

$$\tilde{\mathbf{P}} = \begin{bmatrix} \gamma + (1-\gamma)u_1 & (1-\gamma)(1-u_1) \\ (1-\gamma)u_1 & 1-(1-\gamma)u_1 \end{bmatrix} \tag{227}$$

$$\tilde{\pi} = \begin{bmatrix} u_1 & 1-u_1 \end{bmatrix} \tag{228}$$

For brevity, we let $\mu_n \triangleq \Pr(\exists i, j \in [n], i \neq j, \tilde{H}_i^{(1)} = \tilde{H}_j^{(1)})$. Next, from the union bound, we obtain

$$\mu_n \leq \sum_{(i,j)\in[n]^2 : i<j} \Pr(\tilde{H}_i^{(1)} = \tilde{H}_j^{(1)}) \tag{229}$$

$$\leq n^2 \max_{(i,j)\in[n]^2 : i<j} \Pr(\tilde{H}_i^{(1)} = \tilde{H}_j^{(1)}) \tag{230}$$

Due to stationarity of $\tilde{\mathbf{P}}$, this maximum is equal to $\Pr(\tilde{H}_1^{(1)} = \tilde{H}_{s+1}^{(1)})$ for some s. For brevity, let $\mathbf{Q} \triangleq \tilde{\mathbf{P}}^s$ and $q \triangleq \Pr(\tilde{H}_1^{(1)} = \tilde{H}_{s+1}^{(1)})$. Observe that $\tilde{H}_1^{(1)}$ and $\tilde{H}_{s+1}^{(1)}$ are correlated $\mathrm{Binom}(m_n, 1-u_1)$ random variables and for any s, $\mathbf{Q}$ has positive values, *i.e.,* the collapsed Markov chain is irreducible for any s. Now, we have

$$q = \sum_{r=0}^{m_n} \Pr(\tilde{H}_1^{(1)} = r) \Pr(\tilde{H}_{s+1}^{(1)} = r | \tilde{H}_1^{(1)} = r) \tag{231}$$

$$= \sum_{r=0}^{m_n} \binom{m}{r} (1-u_1)^r u_1^{m_n - r} \Pr(\tilde{H}_{s+1}^{(1)} = r | \tilde{H}_1^{(1)} = r) \tag{232}$$

Note that since the rows of $\tilde{\mathbf{D}}^{(1)}$ are *i.i.d.*, we have

$$\Pr(\tilde{H}_{s+1}^{(1)} = r | \tilde{H}_1^{(1)} = r) = \Pr(M + N = r) \tag{233}$$

where $M \sim \mathrm{Binom}(r, Q_{2,2})$ and $N \sim \mathrm{Binom}(m_n - r, Q_{1,2})$ are independent. Note that there are two ways leading to the state 2 in the collapsed column after s steps. The first one is the state 2 staying in the same state after s steps, and the second one is state 1 being converted to state 2 after s steps. Here the Binomial random variables E and F keep counts of the former and the latter ways, respectively.

Then, from Stirling's approximation [50, Chapter 3.2] on the factorial terms in the Binomial coefficient and [49, Theorem 11.1.2], we get

$$q = \sum_{r=0}^{m_n} \binom{m_n}{r} (1-u_1)^r u_1^{m_n-r} \Pr(M+N=r) \tag{234}$$

$$\leq \frac{e}{\sqrt{2\pi}} {m_n}^{-1/2} \sum_{r=0}^{m_n} \Pi_r^{-1} 2^{-m_n D(\frac{r}{m_n}\|(1-u_1))} \Pr(M+N=r) \tag{235}$$

where $\Pi_r = \frac{r}{m_n}(1-\frac{r}{m_n})$. Let

$$T = \sum_{r=0}^{m_n} \Pi_r^{-1} 2^{-m_n D(\frac{r}{m_n}\|(1-u_1))} \Pr(M+N=r) \tag{236}$$

$$= T_1 + T_2 \tag{237}$$

where

$$T_1 = \sum_{r: D(\frac{r}{m_n}\|1-u_1) > \frac{\varepsilon_n^2}{2\log_e 2}} \Pi_r^{-1} 2^{-m_n D(\frac{r}{m_n}\|(1-u_1))} \Pr(M+N=r) \tag{238}$$

$$T_2 = \sum_{r: D(\frac{r}{m_n}\|1-u_1) \leq \frac{\varepsilon_n^2}{2\log_e 2}} \Pi_r^{-1} 2^{-m_n D(\frac{r}{m_n}\|(1-u_1))} \Pr(M+N=r), \tag{239}$$

$\varepsilon_n > 0$, which is described below in more detail, is such that $\varepsilon_n \to 0$ as $n \to \infty$.

First, we look at T_1. Note that for any $r \in \mathbb{N}$, we have $\Pi_r \leq {m_n}^2$, suggesting the multiplicative term in the summation in (238) is polynomial with m_n. Note that we can simply separate the cases $r = 0$, $r = m_n$ whose probabilities vanish exponentially in m_n. Therefore, as long as $m_n \varepsilon_n^2 \to \infty$, T_1 has a polynomial number of elements which decay exponentially with m_n. Thus

$$T_1 \to 0 \text{ as } n \to \infty \tag{240}$$

as long as $m_n \varepsilon_n^2 \to \infty$.

Now, we focus on T_2. From Pinsker's inequality [49, Lemma 11.6.1], we have

$$D\left(\frac{r}{m_n}\middle\|1-u_1\right) \leq \frac{\varepsilon_n^2}{2\log_e 2} \Rightarrow \mathrm{TV}\left(\frac{r}{m_n}, 1-u_1\right) \leq \varepsilon_n \tag{241}$$

where TV denotes the total variation distance between the Bernoulli distributions with given parameters. Therefore

$$\left|\{r : D\Big(\frac{r}{m_n}\Big\|1-u_1\Big) \leq \frac{\varepsilon_n^2}{2\log_e 2}\}\right| \leq \left|\{r : \mathrm{TV}\Big(\frac{r}{m_n}, 1-u_1\Big) \leq \varepsilon_n\}\right| \tag{242}$$

$$= O(m_n \varepsilon_n) \tag{243}$$

for small ε_n. Furthermore, if $\mathrm{TV}\left(\frac{r}{m_n}, 1-u_1\right) \leq \varepsilon_n$, we have

$$\Pi_r^{-1} \leq \frac{1}{(1-u_1)u_1} \tag{244}$$

Now, we investigate $\Pr(M+N=r)$ for the values of r in the interval $[m_n(1-u_1-\varepsilon_n), m_n(1-u_1+\varepsilon_n)]$.

$$\Pr(M+N=r) = \sum_{i=1}^{r} \Pr(M=r-i)\Pr(N=i) + \Pr(M=r)\Pr(N=0) \tag{245}$$

$$= Q_{2,2}^r Q_{1,1}^{m_n-r} + \sum_{i=1}^{r} \binom{r}{i} Q_{2,2}^{r-i} (1-Q_{2,2})^i \binom{m_n-r}{i} Q_{1,2}^i (1-Q_{1,2})^{m_n-r-i} \tag{246}$$

Again, from Stirling's approximation [50, Chapter 3.2] on the factorial terms in the Binomial coefficient in (246) and from [49, Theorem 11.1.2], we have

$$\Pr(M+N=r) \leq Q_{2,2}^{r} Q_{1,1}^{m_n-r} + \frac{e^2}{2\pi}[r(m_n-r)]^{-1/2} U \tag{247}$$

where

$$U = \sum_{i=1}^{r} \Pi_{i/r}^{-1} \Pi_{i/m_n-r}^{-1} 2^{-rD(1-\frac{i}{r}\|Q_{2,2})-(m_n-r)D(\frac{i}{m_n-r}\|Q_{1,2})} \tag{248}$$

Then, from $r \in [m_n(1-u_1-\varepsilon_n), m_n(1-u_1+\varepsilon_n)]$ we obtain

$$\Pr(M+N=r) \leq Q_{2,2}^{r} Q_{1,1}^{m_n-r} + \frac{e^2}{2\pi} \frac{{m_n}^{-1}}{\sqrt{(1-u_1-\varepsilon_n)(u_1-\varepsilon_n)}} U \tag{249}$$

and

$$U \leq \sum_{i=1}^{r} \Pi_{i/r}^{-1} \Pi_{i/m_n-r}^{-1} 2^{-m_n(1-u_1-\varepsilon_n)D(1-\frac{i}{r}\|Q_{2,2})} 2^{-m_n(u_1-\varepsilon_n)D(\frac{i}{m_n-r}\|Q_{1,2})} \tag{250}$$

$$= \sum_{i \notin \mathscr{R}(\varepsilon_n)} \Pi_{i/r}^{-1} \Pi_{i/m_n-r}^{-1} 2^{-m_n(1-u_1-\varepsilon_n)D(1-\frac{i}{r}\|Q_{2,2})} 2^{-m_n(u_1-\varepsilon_n)D(\frac{i}{m_n-r}\|Q_{1,2})}$$

$$+ \sum_{i \in \mathscr{R}(\varepsilon_n)} \Pi_{i/r}^{-1} \Pi_{i/m_n-r}^{-1} 2^{-m_n(1-u_1-\varepsilon_n)D(1-\frac{i}{r}\|Q_{2,2})} 2^{-m_n(u_1-\varepsilon_n)D(\frac{i}{m_n-r}\|Q_{1,2})} \tag{251}$$

where we define the set $\mathscr{R}(\varepsilon_n)$ as

$$\mathscr{R}(\varepsilon_n) \triangleq \left\{ i \in [r] : D\Big(1-\frac{i}{r}\Big\|Q_{2,2}\Big), D\Big(\frac{i}{m_n-r}\Big\|Q_{1,2}\Big) \leq \frac{\varepsilon_n^2}{2\log_e 2} \right\} \tag{252}$$

Note that similar to T_1, the first summation in (251) vanishes exponentially in m_n whenever $m_n\varepsilon_n^2 \to \infty$, and using Pinsker's inequality once more, the second term can be upper bounded by

$$O(|\mathscr{R}(\varepsilon_n)|) = O(m_n\varepsilon_n) \tag{253}$$

Now, we choose $\varepsilon_n = {m_n}^{-\frac{1}{2}} V_n$ for some V_n satisfying $V_n = \omega(1)$ and $V_n = o(m_n^{1/2})$. Thus, T_1 vanishes exponentially fast since $m_n\varepsilon_n^2 = V_n^2 \to \infty$ and

$$\Pr(M+N=r) = O(\varepsilon_n) \tag{254}$$

$$T = O(m_n\varepsilon_n^2) = O(V_n^2) \tag{255}$$

$$\mu_n = O(n^2 {m_n}^{-1/2} V_n^2) \tag{256}$$

By the assumption $m_n = \omega(n^4)$, we have $m_n = n^4 Z_n$ for some Z_n satisfying $\lim_{n\to\infty} Z_n = \infty$. Now, taking $V_n = o(Z_n^{1/4})$ (e.g. $V_n = Z_n^{1/6}$), we get

$$\mu_n \leq O({Z_n}^{-1/2} V_n^2) = o(1) \tag{257}$$

Thus $m_n = \omega(n^4)$ is sufficient to have $\mu_n \to 0$ as $n \to \infty$, concluding the proof. □

APPENDIX D
PROOF OF ACHIEVABILITY OF THEOREM 3

The proof of the achievability part follows from successive union bounds exploiting the following:

- For any typical row Y^{K_n} of $\mathbf{D}^{(2)}$, there are approximately $2^{K_n H(X|Y)}$ jointly typical sequences with respect to $p_{X,Y}$.
- If the output of the synchronization channel has length K_n then there are at least $k_{\min} = \left\lceil \frac{K_n}{s_{\max}} \right\rceil$ retained (not deleted) elements.
- For the number of columns n, the number of deletion patterns with $k_{\min}$ retained elements is

$$\binom{n}{k_{\min}} \leq 2^{nH_b(k_{\min}/n)} \tag{258}$$

- Any stretched row has the same probability as the original row.
- If the original length-n sequence and the retained length-$k_{\min}$ sequence after the deletion channel are ε-typical with respect to p_X, then the complementary length-$(n-k_{\min})$ subsequence is $\tilde{\varepsilon}$-typical with respect to p_X, where $\tilde{\varepsilon} = \frac{n+k_{\min}}{n-k_{\min}}$.
- The cardinality of the set of $\tilde{\varepsilon}$-typical sequences of length $n-k_{\min}$ with respect to p_X is approximately $2^{(n-k_{\min})H(X)}$.

We need to show that for a given pair of matching rows, WLOG, X_1^n of $\mathbf{D}^{(1)}$ and $Y_l^{K_n}$ of $\mathbf{D}^{(2)}$ with $\Theta_n(1) = l$, the probability of error $P_e \triangleq \Pr(\hat{\Theta}_n(1) \neq l)$ of the following matching scheme can be made arbitrarily small asymptotically where $K_n = \sum_{j=1}^{n} S_{1,j}$ is the random variable corresponding to the length of $Y_l^{K_n}$. The matching scheme we propose follows these steps:

1) For all $j \in [n]$, discard the j^{th} column of $\mathbf{D}^{(1)}$ if $A_j = 1$ to obtain $\bar{\mathbf{D}}^{(1)}$ whose column size is $n-A$ where $A = \sum_{j=1}^{n} A_j$.

2) Stretch each row $\bar{X}_i^{n-A} = \bar{X}_{i,1}, \ldots, \bar{X}_{i,n-A}$ of $\bar{\mathbf{D}}^{(1)}$ into $\tilde{X}_i^{(n-A)s_{\max}}$, by repeating each element of $\bar{X}_i^{n-A}$ $s_{\max}$ times as follows

$$\tilde{X}_i^{(n-A)s_{\max}} = 1^{s_{\max}} \otimes \bar{X}_{i,1}, \ldots, 1^{s_{\max}} \otimes \bar{X}_{i,n-A} \tag{259}$$

where $1^{s_{\max}}$ is an all-one row vector of length $s_{\max}$ and $\otimes$ denotes the Kronecker product.

3) Fix $\varepsilon > 0$. If $K_n < k \triangleq n(\mathbb{E}[S] - \varepsilon)$ declare error, whose probability is denoted by κ_n where k is assumed to be an integer for computational simplicity. Otherwise, proceed with the next step.

4) If $A < a = n(\alpha\delta - \varepsilon)$ declare error, whose probability is denoted by μ_n. Otherwise, proceed with the next step.

5) Match the l^{th} row $Y_l^{K_n}$ of $\mathbf{D}^{(2)}$ X_1^n of $\mathbf{D}^{(1)}$, assigning $\hat{\Theta}_n(1) = l$, if $i = 1$ is the only index in $[m_n]$ such that *i)* $\bar{X}_i^{n-A}$ is ε-typical and *ii)* $\tilde{X}_i^{(n-A)s_{\max}}$ contains a subsequence jointly ε-typical with $Y_l^{K_n}$ with respect to $p_{X,Y}$. Otherwise, declare a *collision* error.

Since additional columns in $\mathbf{D}^{(2)}$ and additional detected deleted columns in $\mathbf{D}^{(1)}$ would decrease the collision probability, we have

$$\Pr(\text{collision between 1 and } i | K_n \geq k, A \geq a) \leq \Pr(\text{collision between 1 and } i || K_n = k, A = a) \tag{260}$$

for any $i \in [m_n] \setminus \{1\}$. Thus, we can focus on the case $K_n = k$, $A = a$, as it yields an upper bound on the error probability of our matching scheme.

Let $A_\varepsilon^{(n-a)}(X)$ denote the set of ε-typical (with respect to p_X) sequences of length $n-a$ and $A_\varepsilon(X^k|Y_l^k)$ denote the set of sequences of length k jointly ε-typical (with respect to $p_{X,Y}$) with Y_l^k. For the matching rows X_1^n, Y_l^k of $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$, define the pairwise collision probability between X_1^n and X_i^n for any $i \in [m_n] \setminus \{1\}$ as

$$P_{\text{col,i}} \triangleq \Pr(\exists z^k : z^k \in A_\varepsilon(X^k|Y_l^k) \text{ and } z^k \text{ is a subsequence of } \tilde{X}_i^{(n-a)s_{\max}}.). \tag{261}$$

Therefore given the correct labeling for $Y_l^k \in \mathbf{D}^{(2)}$ is $X_1^n \in \mathbf{D}^{(1)}$, the probability of error P_e can be bounded as

$$\begin{aligned} P_e &\leq \Pr(\nexists z^k : z^k \in A_\varepsilon(X^k|Y_l^k) \text{ and } z^k \text{ is a subsequence of } \tilde{X}_1^{(n-a)s_{\max}}.) \\ &\quad + \Pr(X_1^n \notin A_\varepsilon^{(n)}(X)) + \sum_{i=2}^{2^{nR}} P_{\text{col,i}} + \kappa_n + \mu_n \qquad (262) \\ &\leq 2\varepsilon + \sum_{i=2}^{2^{nR}} P_{\text{col,i}} + \kappa_n + \mu_n \qquad (263) \\ &\leq 2\varepsilon + 2^{nR} P_{\text{col,2}} + \kappa_n + \mu_n \qquad (264) \end{aligned}$$

where (264) follows from the fact the the rows are *i.i.d.* and thus $P_{\text{col,i}} = P_{\text{col,2}}$, $\forall i \in [m_n] \setminus \{1\}$.

We now upper bound $P_{\text{col,2}}$. First, we investigate repetition distributions with $\frac{1}{s_{\max}}\mathbb{E}[S] \geq \frac{1-\alpha\delta}{|\mathfrak{X}|}$. Let $F(n,k,|\mathfrak{X}|)$ denote the number of $|\mathfrak{X}|$-ary sequences of length n, which contain a fixed $|\mathfrak{X}|$-ary sequence of length k. We note that this $F(n,k,|\mathfrak{X}|)$ is constant for any $|\mathfrak{X}|$-ary sequence of length k [63, Lemma 1]. Now we define $G_{z^k}(ns_{\max},k,|\mathfrak{X}|)$ as the number of $s_{\max}$ times stretched sequences of length $ns_{\max}$, containing a $|\mathfrak{X}|$-ary sequence z^k of length k. We stress that this counting function G_{z^k} will not be independent of z^k as is the case for the counting function F. For example, let $s_{\max} = 2$, $\mathfrak{X} = \{0,1\}$, $n = 2$, $k = 2$, $z_1^k = 01$ and $z_2^k = 00$. Then we have $G_{z_1^k}(ns_{\max},k,|\mathfrak{X}|) = 1$ since only 0011 contains $z_1^k = 01$, whereas $G_{z_2^k}(ns_{\max},k,|\mathfrak{X}|) = 3$ since 0000, 0011 and 1100 all contain $z_2^k = 00$.

Observe that the maximum value of $G_{z^k}(ns_{\max},k,|\mathfrak{X}|)$ is attained when z^k consists only of one symbol repeated k times, as this grouping of elements in z^k yields the maximum number of possible elementwise replicated sequences. WLOG, let $z^k = 00\ldots0$. Then, to count $G_{z^k}(ns_{\max},k,|\mathfrak{X}|)$, we group the consecutive $s_{\max}$ 0's in z^k together, allowing the last group to have possibly fewer than $s_{\max}$ elements. It is clear that there are $\left\lceil \frac{k}{s_{\max}} \right\rceil$ of such groups of 0's. Since we put a stretching constraint on the sequences of length $ns_{\max}$ when we count $G_{z^k}(ns_{\max},k,|\mathfrak{X}|)$, we are looking for sequences of length n, containing a subsequence of length $\left\lceil \frac{k}{s_{\max}} \right\rceil$. Thus, counting this number will be the same as counting $F\left(n, \left\lceil \frac{k}{s_{\max}} \right\rceil, |\mathfrak{X}|\right)$. Thus we have

$$G_{z^k}(ns_{\max},k,|\mathfrak{X}|) \leq F\left(n, \lceil k/s_{\max} \rceil, |\mathfrak{X}|\right), \quad \forall z^k \in \mathfrak{X}^k \qquad (265)$$

We note that the inequality given in (265) is the tightest upper bound independent of z^k, equality being achieved when z^k is a constant (*e.g.,* all-zeros) sequence.

Now, let

$$T(z^k,A^n) \triangleq \{x^n \in \mathfrak{X}^n : \bar{x}^{(n-a)} \in A_\varepsilon^{(n-a)}(X) \text{ and } \tilde{x}^{(n-a)s_{\max}} \text{ contains } z^k.\} \qquad (266)$$

Then, we obtain

$$\begin{aligned} |T(z^k,A^n)| &\leq G_{z^k}((n-a)s_{\max},k,|\mathfrak{X}|) \qquad (267) \\ &\leq F\left(n-a, \lceil k/s_{\max} \rceil, |\mathfrak{X}|\right) \qquad (268) \end{aligned}$$

For the sake of computational simplicity, suppose $\frac{k}{s_{\max}}$ is an integer. Since $\frac{1}{s_{\max}}\mathbb{E}[S] \geq \frac{1-\alpha\delta}{|\mathfrak{X}|}$, from [63] and [49, Chapter 11] we have the following upper bound:

$$F\left(n-a, k/s_{\max}, |\mathfrak{X}|\right) \leq (n-a)2^{(n-a)H_b\left(\frac{k}{s_{\max}(n-a)}\right)}(|\mathfrak{X}|-1)^{\left(n-a-\frac{k}{s_{max}}\right)} \qquad (269)$$

Furthermore, for any $x^n \in T(z^k,A^n)$, since $T(z^k,A^n) \subseteq A_\varepsilon^{(n-a)}(X)$, we have

$$p_{X^n}(x^n) \leq 2^{-(n-a)(H(X)-\varepsilon)} \qquad (270)$$

and since the rows X_i^n of $\mathbf{D}^{(1)}$ are *i.i.d.*, we have

$$\Pr(X_2^n \in T(z^k,A^n)|X_1^n \in T(z^k,A^n)) = \Pr(X_2^n \in T(z^k,A^n)) \tag{271}$$

Finally, we have

$$|A_\varepsilon(X^k|Y_l^k)| \leq 2^{k(H(X|Y)+\varepsilon)} \tag{272}$$

Combining (268)-(272), we can upper bound $P_{\text{col},2}$ as

$$\begin{aligned}
P_{\text{col},2} &\leq \sum_{z^k \in A_\varepsilon(X^k|Y_l^k)} \Pr(X_2^n \in T(z^k,A^n)) && (273)\\
&= \sum_{z^k \in A_\varepsilon(X^k|Y_l^k)} \sum_{x^n \in T(z^k,A^n)} p_{X^n}(x^n) && (274)\\
&\leq \sum_{z^k \in A_\varepsilon(X^k|Y_l^k)} \sum_{x^n \in T(z^k,A^n)} 2^{-(n-a)(H(X)-\varepsilon)} && (275)\\
&= \sum_{z^k \in A_\varepsilon(X^k|Y_l^k)} |T(z^k,A^n)| 2^{-(n-a)(H(X)-\varepsilon)} && (276)\\
&\leq \sum_{z^k \in A_\varepsilon(X^k|Y_l^k)} 2^{-(n-a)(H(X)-\varepsilon)} F\left(n-a, k/s_{\max}, |\mathfrak{X}|\right) && (277)\\
&= |A_\varepsilon(X^k|Y_l^k)| 2^{-(n-a)(H(X)-\varepsilon)} F\left(n-a, k/s_{\max}, |\mathfrak{X}|\right) && (278)\\
&\leq |A_\varepsilon(X^k|Y_l^k)| (n-a) 2^{-(n-a)\left[H(X)-\varepsilon-H_b\left(\frac{k}{s_{\max}(n-a)}\right)\right]} (|\mathfrak{X}|-1)^{\left(n-a-\frac{k}{s_{\max}}\right)} && (279)\\
&\leq 2^{k(H(X|Y)+\varepsilon)} (n-a) 2^{-(n-a)\left[H(X)-\varepsilon-H_b\left(\frac{k}{s_{\max}(n-a)}\right)\right]} (|\mathfrak{X}|-1)^{\left(n-a-\frac{k}{s_{max}}\right)} && (280)
\end{aligned}$$

Thus, we have the following upper bound on the error probability

$$P_e \leq 2\varepsilon + 2^{nR} 2^{k(H(X|Y)+\varepsilon)} (n-a) 2^{-(n-a)\left[H(X)-\varepsilon-H_b\left(\frac{k}{s_{\max}(n-a)}\right)\right]} (|\mathfrak{X}|-1)^{\left(n-a-\frac{k}{s_{max}}\right)} + \kappa_n + \mu_n \tag{281}$$

By LLN, we have $\kappa_n \to 0$ and $\mu_n \to 0$ as $n \to \infty$. Hence, we can argue that any database growth rate R satisfying

$$R < \left[(1-\alpha\delta)\left(H(X) - H_b\left(\frac{\mathbb{E}[S]}{(1-\alpha\delta)s_{\max}}\right)\right) - \left(1-\alpha\delta-\frac{\mathbb{E}[S]}{s_{\max}}\right)\log\left(|\mathfrak{X}|-1\right) - \mathbb{E}[S]H(X|Y)\right]^+ \tag{282}$$

is achievable, by taking ε small enough.

Now, we focus on general repetition distributions. For any subsequence z^k of $s_{\max}$-times stretched sequence of length $(n-a)s_{\max}$, let $r(z^k)$ be the number of runs in z^k with at most $s_{\max}$ elements and note that $r(z^k) \leq n-a$. Then, let $\tilde{z}^{r(z^k)}$ be the sequence storing the values of each run in z^k. Observe that for any $z^k \in A_\varepsilon(X^k|Y_l^k)$, we have $\tilde{z}^{r(z^k)} \in A_\varepsilon^{(r(z^k))}(X)$.

For any such grouping of $r(z^k)$ runs, the ε-typicality of $x^n = (x_1,\ldots,x_n) \in T(z^k,A^n)$ and $\tilde{z}^{r(z^k)}$ with respect to p_X implies the $\tilde{\varepsilon}$-typicality of the remaining sequence of length $n-a-r(z^k)$ obtained after discarding $\tilde{z}^{r(z^k)}$ from $\vec{x}^{n-a}$, where $\tilde{\varepsilon} = \frac{n-a+r(z^k)}{n-a-r(z^k)}\varepsilon$. Furthermore, by a similar argument made above, we stress that $T(z^k,A^n)$ attains its maximum value when $r(z^k)$ is the minimum, which is $k_{\min} \triangleq \lceil \frac{k}{s_{\max}} \rceil$, attained

when z^k is a $s_{\max}$ times stretched sequence itself. Therefore for any $z^k \in A_\varepsilon(X^k|Y_l^k)$, taking the union bound over all possible groupings with $r(z^k)$ runs, the cardinality of $T(z^k,A^n)$ can be upper bounded as

$$|T(z^k,A^n)| \leq \binom{n-a}{k_{\min}} |A_{\tilde{\varepsilon}}^{(n-a-k_{\min})}(X)| \tag{283}$$

$$\leq 2^{(n-a)H_b\left(\frac{k_{\min}}{n-a}\right)} |A_{\tilde{\varepsilon}}^{(n-a-\hat{k})}(X)| \tag{284}$$

$$\leq 2^{(n-a)H_b\left(\frac{k_{\min}}{n-a}\right)} 2^{(n-a-k_{\min})(H(X)+\tilde{\varepsilon})} \tag{285}$$

$$= 2^{n\left[(1-\frac{a}{n})H_b\left(\frac{k_{\min}}{n-a}\right)+(1-\frac{a}{n}-\frac{k_{\min}}{n})(H(X)+\tilde{\varepsilon})\right]} \tag{286}$$

Plugging (286) into (276) and following the same steps, one can show that any rate R satisfying

$$R < \left[\frac{\mathbb{E}[S]}{s_{\max}} H(X) - (1-\alpha\delta) H_b\left(\frac{\mathbb{E}[S]}{(1-\alpha\delta)s_{\max}}\right) - \mathbb{E}[S]H(X|Y)\right]^+ \tag{287}$$

is achievable. Simply taking the maximum of the two proven achievable rates ((282) and (287)) when $\frac{1}{s_{\max}}\mathbb{E}[S] \geq \frac{1-\alpha\delta}{|\mathfrak{X}|}$ yields (67). This concludes the proof. □

Appendix E
Proof of Corollary 3

Let E denote the empty string and $\tilde{X}$ denote the sequence obtained after discarding the detected deleted entries from X^2. The dependence of $\tilde{X}$ on X^2 and A^2 and that of Y on X^2 and S^2 are omitted for brevity.

We start with the fact that since the entries of X^2 are independent, the deleted entries do not offer any information. Thus, we can discard them without any information loss. Thus, we have

$$I(X^2;Y,A^2) = I(\tilde{X};Y|A^2) \tag{288}$$

$$= H(\tilde{X}|A^2) - H(\tilde{X}|Y,A^2) \tag{289}$$

We have

$$H(\tilde{X}|A^2) = \sum_{a^2\in\{0,1\}^2} \Pr(A^2=a^2) H(\tilde{X}|A^2=a^2) \tag{290}$$

$$= \Pr(A^2=00)H(\tilde{X}|A^2=00) + \Pr(A^2=01)H(\tilde{X}|A^2=01) + \Pr(A^2=10)H(\tilde{X}|A^2=10) + \Pr(A^2=11)H(\tilde{X}|A^2=11) \tag{291}$$

$$= (1-\alpha\delta)^2 2H(X) + \alpha\delta(1-\alpha\delta)H(X) + (1-\alpha\delta)\alpha\delta H(X) + 0 \tag{292}$$

$$= 2(1-\alpha\delta)H(X) \tag{293}$$

Furthermore, we have

$$H(\tilde{X}|Y,A^2) = \sum_{y,a^2} \Pr(Y=y,A^2=a^2) H(\tilde{X}|Y=y,A^2=a^2) \tag{294}$$

$$\begin{aligned} &= \Pr(Y=E,A^2=00)H(\tilde{X}|Y=E,A^2=00) \\ &\quad + \Pr(Y=E,A^2=01)H(\tilde{X}|Y=E,A^2=01) \\ &\quad + \Pr(Y=E,A^2=10)H(\tilde{X}|Y=E,A^2=10) \\ &\quad + \sum_{x\in\mathfrak{X}} \Pr(Y=x,A^2=00)H(\tilde{X}|Y=x,A^2=00) \\ &\quad + \sum_{x\in\mathfrak{X}} \Pr(Y=x,A^2=01)H(\tilde{X}|Y=x,A^2=01) \\ &\quad + \sum_{x\in\mathfrak{X}} \Pr(Y=x,A^2=10)H(\tilde{X}|Y=x,A^2=10) \end{aligned} \tag{295}$$

Note that in (295), we discarded the terms with $A^2 = 11$ for $|Y| \geq 1$, since $\Pr(|Y| \geq 1, A^2 = 11) = 0$. We can further discard the terms with $|Y| = n = 2$, since in that case we have no deletion and $Y = Y^2 = X^2$. Finally, we can also discard the last two terms in (295) since for any $x \in \mathfrak{X}$ we have

$$H(\tilde{X}|Y = x, A^2 = 01) = H(\tilde{X}|Y = x, A^2 = 10) = 0 \tag{296}$$

Thus, we have

$$\begin{aligned} H(\tilde{X}|Y, A^2) &= \delta^2(1-\alpha)^2 2H(X) + \delta^2(1-\alpha)\alpha H(X) + \delta^2\alpha(1-\alpha)H(X) \\ &\quad + \sum_{x\in\mathfrak{X}} \Pr(Y = x, A^2 = 00) H(\tilde{X}|Y = x, A^2 = 00) \qquad (297) \\ &= 2\delta^2(1-\alpha)H(X) + \sum_{x\in\mathfrak{X}} \Pr(Y = x, A^2 = 00) H(\tilde{X}|Y = x, A^2 = 00) \qquad (298) \end{aligned}$$

We first compute $\Pr(Y = x, A^2 = 00)$. For any $x \in \mathfrak{X}$, we have

$$\begin{aligned} \Pr(Y = x, A^2 = 00) &= \sum_{x^2\in\mathfrak{X}^2} \Pr(Y = x, A^2 = 00, X^2 = x^2) \qquad (299) \\ &= \Pr(Y = x, A^2 = 00, X^2 = xx) + 2\sum_{y\neq x} \Pr(Y = x, A^2 = 00, X^2 = xy) \qquad (300) \\ &= p_X(x)^2 2\delta(1-\delta)(1-\alpha) + 2\sum_{y\neq x} p_X(x)p_X(y)\delta(1-\delta)(1-\alpha) \qquad (301) \\ &= 2\delta(1-\delta)(1-\alpha)p_X(x)\sum_{y\in\mathfrak{X}} p_X(y) \qquad (302) \\ &= 2\delta(1-\delta)(1-\alpha)p_X(x) \qquad (303) \end{aligned}$$

Now, we compute $H(\tilde{X}|Y = x, A^2 = 00)$. For any $x \in \mathfrak{X}$ we have $2|\mathfrak{X}| - 1$ possible patterns for $\tilde{X}$, given that $Y = x$. $2|\mathfrak{X}| - 2$ of these patterns have probabilities proportional to $p_X(x)p_X(y)$ $y \in \mathfrak{X} \setminus \{x\}$ and the remaining pattern has probability proportional to $2p_X(x)^2$. Thus we have

$$H(\tilde{X}|Y = x, A^2 = 00) = H\Big(\frac{p_X(1)p_X(x)}{c}, \frac{p_X(x)p_X(1)}{c}, \ldots, \frac{2(p_X(x))^2}{c}, \ldots, \frac{p_X(|\mathfrak{X}|)p_X(x)}{c}, \frac{p_X(x)p_X(|\mathfrak{X}|)}{c}\Big) \tag{304}$$

where the normalization constant c is $c = 2p_X(x)$. Thus,

$$\begin{aligned} H(\tilde{X}|Y = x, A^2 = 00) &= H\Big(\frac{p_X(1)}{2}, \frac{p_X(1)}{2}, \ldots, p_X(x), \ldots, \frac{p_X(|\mathfrak{X}|)}{2}, \frac{p_X(|\mathfrak{X}|)}{2}\Big) \qquad (305) \\ &= H(X) + 1 - p_X(x) \qquad (306) \end{aligned}$$

Combining (298)-(306), we can compute $H(\tilde{X}|Y, A^2)$ as

$$\begin{aligned} H(\tilde{X}|Y, A^2) &= 2\delta^2(1-\alpha)H(X) + \sum_{x\in\mathfrak{X}} 2\delta(1-\delta)(1-\alpha)p_X(x)[H(X) + 1 - p_X(x)] \qquad (307) \\ &= 2\delta^2(1-\alpha)H(X) + 2\delta(1-\delta)(1-\alpha)\left(H(X) + 1 - \hat{q}\right) \qquad (308) \\ &= 2\delta(1-\alpha)H(X) + 2\delta(1-\delta)(1-\alpha)\left(1 - \hat{q}\right) \qquad (309) \end{aligned}$$

Finally, combining (293) and (309), we obtain

$$\begin{aligned} I(\tilde{X}; Y^K|A^2) &= H(\tilde{X}|A^2) - H(\tilde{X}|Y(X^2), A^2) \qquad (310) \\ &= 2(1-\alpha\delta)H(X) - 2\delta(1-\alpha)H(X) - 2\delta(1-\delta)(1-\alpha)\left(1 - \hat{q}\right) \qquad (311) \\ &= 2(1-\delta)H(X) - 2\delta(1-\delta)(1-\alpha)\left(1 - \hat{q}\right) \qquad (312) \end{aligned}$$

Thus, we have

$$\frac{1}{2}I(X^2; Y^K, A^2) = (1-\delta)H(X) - \delta(1-\delta)(1-\alpha)\left(1 - \hat{q}\right) \tag{313}$$

concluding the proof. □

APPENDIX F
PROOF OF COROLLARY 4

We start by observing that

$$I(X^2;Y,A^2) = I(X^2;Y,|Y|,A^2) \tag{314}$$
$$= H(X^2) - H(X^2|Y,|Y|,A^2) \tag{315}$$
$$= 2H(X) - H(X^2|Y,|Y|,A^2) \tag{316}$$

Furthermore, we have

$$H(X^2|Y,|Y|,A^2) = \sum_{i=0}^{2} \Pr(|Y|=i) H(X^2|Y,|Y|=i,A^2) \tag{317}$$
$$\begin{aligned} &= \delta^2 H(X^2|Y,|Y|=0,A^2) \\ &\quad + 2\delta(1-\delta) H(X^2|Y,|Y|=1,A^2) \\ &\quad + (1-\delta)^2 H(X^2|Y,|Y|=2,A^2) \end{aligned} \tag{318}$$
$$\begin{aligned} &= \delta^2 2H(X) \\ &\quad + 2\delta(1-\delta) H(X^2|Y,|Y|=1,A^2) \\ &\quad + (1-\delta)^2 2H(X|Y) \end{aligned} \tag{319}$$
$$\begin{aligned} &= \delta^2 2H(X) \\ &\quad + 2\delta(1-\delta)\alpha[H(X)+H(X|Y)] \\ &\quad + 2\delta(1-\delta)(1-\alpha) H(X^2|Y,|Y|=1,A^2=00) \\ &\quad + (1-\delta)^2 2H(X|Y) \end{aligned} \tag{320}$$

Note that we can rewrite $H(X^2|Y,|Y|=1,A^2=00)$ as

$$H(X^2|Y,|Y|=1,A^2=00) = H(X^2|Y,|Y|=1) \tag{321}$$
$$= 2H(X) - I(X^2;Y||Y|=1) \tag{322}$$
$$= 2H(X) - [H(Y) - H(Y|X^2,|Y|=1)] \tag{323}$$

where we have

$$H(Y|X^2,|Y|=1) = \sum_{x^2 \in \mathcal{X}^2} \Pr(X^2=x^2) H(Y|X^2=x^2,|Y|=1) \tag{324}$$

Writing the sum in (324) explicitly, we obtain

$$\begin{aligned} H(Y|X^2,|Y|=1) &= (1-p)^2 H(Y|X^2=00,|Y|=1) + p^2 H(Y|X^2=11,|Y|=1) \\ &\quad + p(1-p) H(Y|X^2=01,|Y|=1) + p(1-p) H(Y|X^2=10,|Y|=1) \end{aligned} \tag{325}$$

Observing the following,

$$H(Y|X^2=00,|Y|=1) = H(Y|X=0) \tag{326}$$
$$H(Y|X^2=11,|Y|=1) = H(Y|X=1) \tag{327}$$
$$H(Y|X^2=01,|Y|=1) = H(V) \tag{328}$$
$$H(Y|X^2=10,|Y|=1) = H(V) \tag{329}$$
$$H(Y|X=0) + H(Y|X=1) = 2\left[\frac{1}{2}H(Y|X=0) + \frac{1}{2}H(Y|X=1)\right] \tag{330}$$
$$= 2H(V|U) \tag{331}$$

we obtain

$$\begin{aligned} H(Y|X^2,|Y|=1) &= (1-p)H(Y|X=0) - p(1-p)H(Y|X=0) \\ &\quad + pH(Y|X=1) - p(1-p)H(Y|X=1) \\ &\quad + 2p(1-p)H(V) \qquad (332) \\ &= H(Y|X) + 2p(1-p)I(U;V) \qquad (333) \end{aligned}$$

Hence, we have

$$H(X^2|Y,|Y|=1,A^2=00) = 2H(X) - I(X;Y) + 2p(1-p)I(U;V) \tag{334}$$

Combining (316)-(334), have

$$\frac{1}{2}I(X^2;Y^K,A^2) = (1-\delta)I(X;Y) - 2\delta(1-\delta)(1-\alpha)p(1-p)I(U;V) \tag{335}$$

concluding the proof. □

APPENDIX G
PROOF OF THEOREM 4

First, we focus on $\delta \leq 1-\hat{q}$ and prove the achievability part. For a given pair of matching rows, WLOG, X_1^n of $\mathbf{D}^{(1)}$ and $Y_l^{K_n}$ of $\mathbf{D}^{(2)}$ with $\Theta_n(1)=l$, let $P_e \triangleq \Pr(\hat{\Theta}_n(1)\neq l)$ be the probability of error of the following matching scheme:

1) Construct the collapsed histogram vectors $\tilde{H}_j^{(1),n}$ and $\tilde{H}_j^{(2),K_n}$ as

$$\tilde{H}_j^{(r)} = \sum_{i=1}^{m_n} \mathbb{1}_{\left[\tilde{D}_{i,j}^{(r)}=2\right]}, \quad \begin{cases} \forall j \in [n], & \text{if } r=1 \\ \forall j \in [K_n] & \text{if } r=2 \end{cases} \tag{336}$$

where K_n denotes the column size of $\mathbf{D}^{(2)}$.

2) Check the uniqueness of the entries $\tilde{H}_j^{(1)}$ $j\in[n]$ of $\tilde{H}^{(1),n}$. If there are at least two which are identical, declare a *detection error* whose probability is denoted by μ_n. Otherwise, proceed with Step 3.

3) $\forall i \in [n]$ if $\nexists j \in [K_n]$, $\tilde{H}_i^{(1)} = \tilde{H}_j^{(2)}$, declare the i^{th} column of $\mathbf{D}^{(1)}$ deleted, assigning $i \in \hat{I}_{\text{del}}$. Note that conditioned on Step 2, this step is error-free.

4) Match the l^{th} row $Y_l^{K_n}$ of $\mathbf{D}^{(2)}$ with the 1^{st} row X_1^n of $\mathbf{D}^{(1)}$, assigning $\hat{\Theta}_n(1)=l$ if the 1^{st} row $\hat{X}_1^{K_n}(\hat{I}_{\text{del}})$ of $\hat{\mathbf{D}}^{(1)}$ is the only row of $\hat{\mathbf{D}}^{(1)}$ equal to $Y_l^{K_n}$ where $\hat{X}_i^{K_n}(\hat{I}_{\text{del}})$ is obtained by discarding the elements of X_i^n whose indices lie in $\hat{I}_{\text{del}}$. Otherwise, declare a *collision error*.

Let $I(\delta)$ be the set of all deletion patterns with up to $n\delta$ deletions. For the matching rows X_1^n, Y_l^k of $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$, define the pairwise adversarial collision probability between X_1^n and X_i^n for any $i \in [m_n]\setminus\{1\}$ as

$$\begin{aligned} P_{\text{col,i}} &\triangleq \Pr(\exists \hat{I}_{\text{del}} \in I(\delta) : \hat{X}_i^{K_n}(\hat{I}_{\text{del}}) = Y_l^{K_n}) \qquad (337) \\ &= \Pr(\exists \hat{I}_{\text{del}} \in I(\delta) : \hat{X}_i^{K_n}(\hat{I}_{\text{del}}) = \hat{X}_1^{K_n}(\hat{I}_{\text{del}})). \qquad (338) \end{aligned}$$

Note that the statement $\exists \hat{I}_{\text{del}} \in I(\delta) : \hat{X}_i^{K_n}(\hat{I}_{\text{del}}) = \hat{X}_1^{K_n}(\hat{I}_{\text{del}})$ is equivalent to the case when the Hamming distance between X_i^n and X_1^n being upper bounded by $n\delta$. In other words,

$$P_{\text{col,i}} = \Pr(d_H(X_1^n, X_i^n) \leq n\delta) \tag{339}$$

where

$$d_H(X_1^n, X_i^n) = \sum_{j=1}^{n} \mathbb{1}_{[X_{1,j}\neq X_{i,j}]} \tag{340}$$

Note that due to the *i.i.d.* nature of the database elements, $d_H(X_1^n, X_i^n) \sim \text{Binom}(n, 1-\hat{q})$. Thus, for any $\delta \leq 1-\hat{q}$, using Chernoff bound [61, Lemma 4.7.2], we have

$$P_{\text{col,i}} = \Pr(d_H(X_1^n, X_i^n) \leq n\delta) \tag{341}$$
$$\leq 2^{-nD(\delta\|1-\hat{q})} \tag{342}$$

Therefore given the correct labeling for $Y_l^k \in \mathbf{D}^{(2)}$ is $X_1^n \in \mathbf{D}^{(1)}$, the probability of error P_e can be bounded as

$$P_e \leq \Pr(\exists i \in [m_n] \setminus \{1\} : \hat{X}_i^{K_n} = \hat{X}_1^{K_n}) \tag{343}$$
$$\leq \sum_{i=2}^{2^{nR}} P_{col,i} + \kappa_n \tag{344}$$
$$\leq 2^{nR} P_{\text{col,2}} + \kappa_n \tag{345}$$

where (345) follows from the fact the the rows are *i.i.d.* and thus $P_{\text{col,i}} = P_{\text{col,2}}, \forall i \in [m_n] \setminus \{1\}$. Combining (342)-(345), we get

$$P_e \leq 2^{nR} \Pr(d_H(X_1^n, X_i^n) \leq n\delta) + \kappa_n \tag{346}$$
$$\leq 2^{nR} 2^{-nD(\delta\|1-\hat{q})} + \kappa_n \tag{347}$$
$$= 2^{-n[D(\delta\|1-\hat{q})-R]} + \kappa_n \tag{348}$$

By Lemma 3, $\kappa_n \to 0$ as $n \to \infty$. Thus, we argue that any rate R satisfying

$$R < D(\delta\|1-\hat{q}) \tag{349}$$

is achievable.

Now we prove the converse part. Suppose $P_e \to 0$. Then, we have

$$P_e = \Pr(\exists i \in [m_n] \setminus \{1\} : d_H(X_1^n, X_i^n) \leq n\delta) \tag{350}$$
$$= 1 - \Pr(\forall i \in [m_n] \setminus \{1\} : d_H(X_1^n, X_i^n) > n\delta) \tag{351}$$
$$= 1 - \prod_{i=2}^{m_n} \Pr(d_H(X_1^n, X_i^n) > n\delta) \tag{352}$$
$$= 1 - \prod_{i=2}^{m_n} [1 - \Pr(d_H(X_1^n, X_i^n) \leq n\delta)] \tag{353}$$
$$= 1 - [1 - \Pr(d_H(X_1^n, X_2^n) \leq n\delta)]^{m_n - 1} \tag{354}$$

where (351)-(354) follow from the *i.i.d.*ness of the rows of $\mathbf{D}^{(1)}$. Since $D_{n,2} \sim \text{Binom}(n, 1-\hat{q})$, for $\delta \leq 1-\hat{q}$, from [61, Lemma 4.7.2], we obtain

$$\Pr(D_{n,2} \leq n\delta) \geq \frac{2^{-nD(\delta\|1-\hat{q})}}{\sqrt{2n}} \tag{355}$$

Plugging (355) into (354), we get

$$P_e \geq 1 - \left[1 - \frac{2^{-nD(\delta\|1-\hat{q})}}{\sqrt{2n}}\right]^{m_n - 1} \tag{356}$$

Now let $y = -\frac{2^{-nD(\delta\|1-\hat{q})}}{\sqrt{2n}} \in (-1, 0)$. Then, we get

$$P_e \geq 1 - (1+y)^{m_n - 1} \tag{357}$$

Since $y \geq -1$, and $m_n \in \mathbb{N}$, we have

$$1 + y(m_n - 1) \leq (1+y)^{m_n - 1} \leq e^{y(m_n-1)} \tag{358}$$

where the LHS of (358) follows from Bernoulli's inequality [64, Theorem 1] and the RHS of (358) follows from the fact that

$$\forall x \in \mathbb{R}, \quad \forall r \in \mathbb{R}_{\geq 0} \quad (1+x)^r \leq e^{xr} \tag{359}$$

Thus, we get

$$P_e \geq 1 - (1+y)^{m_n - 1} \tag{360}$$

$$\geq 1 - e^{y(m_n-1)} \tag{361}$$

$$\geq 0 \tag{362}$$

since $y < 0$, $m_n - 1 > 0$. Note that since $P_e \to 0$, by the Squeeze Theorem [64, Theorem 2], we have

$$\lim_{n\to\infty} 1 - e^{y(m_n-1)} \to 0. \tag{363}$$

This, in turn, implies $ym_n \to 0$ since the exponential function is continuous everywhere. In other words,

$$\lim_{n\to\infty} -\frac{2^{-nD(\delta\|1-\hat{q})}}{\sqrt{2n}} m_n \to 0. \tag{364}$$

Equivalently, from the continuity of the logarithm function, we get

$$\lim_{n\to\infty} -nD(\delta\|1-\hat{q}) + \log m_n - \frac{1}{2}\log(2n) \to -\infty \tag{365}$$

$$\lim_{n\to\infty} -n\left[D(\delta\|1-\hat{q}) - \frac{1}{n}\log m_n + \frac{\log(2n)}{2n}\right] \to -\infty \tag{366}$$

$$\lim_{n\to\infty} \left[D(\delta\|1-\hat{q}) - \frac{1}{n}\log m_n + \frac{\log(2n)}{2n}\right] \geq 0 \tag{367}$$

This implies

$$D(\delta\|1-\hat{q}) \geq \lim_{n\to\infty} \frac{1}{n}\log m_n \tag{368}$$

$$= R \tag{369}$$

finishing the proof for $\delta \leq 1 - \hat{q}$. Thus, we have showed that

$$C^{\text{adv}}(\delta) = D(\delta\|1-\hat{q}) \tag{370}$$

for $\delta \leq 1 - \hat{q}$.

We argue that for $\delta > 1 - \hat{q}$, the adversarial matching capacity is zero, by using two facts: *i)* Since the adversarial deletion budget is an upper bound on deletions, the adversarial matching capacity satisfies

$$C^{\text{adv}}(\delta) \leq C^{\text{adv}}(\delta'), \quad \forall \delta' \leq \delta \tag{371}$$

and *ii)* $C^{\text{adv}}(1-\hat{q}) = 0$. Thus, $\forall \delta > 1 - \hat{q}$, $C^{\text{adv}}(\delta) = 0$. This finishes the proof. □

APPENDIX H
PROOF OF THEOREM 5

First, note that the converse part of Theorem 5 (equation (93)) is trivially true since $C(\infty,0)$ is a non-decreasing function of the seed size Λ_n. Hence it is sufficient to prove the achievability part of Theorem 5 (equation (92)).

For the achievability, we use a matching scheme which *i)* utilizes replica detection and marker addition as done in Section III-C and *ii)* checks the existence of jointly typical subsequences as done in Section IV-A. The matching scheme we propose is as follows:

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

2) Based on the replica detection step, place markers between the noisy replica runs of different columns to obtain $\tilde{\mathbf{D}}^{(2)}$. Note that at this step we cannot detect runs of length 0 as done in Section III-C. Therefore conditioned on the success of the replica detection we have $\tilde{K}_n = \sum_{j=1}^{n} \mathbb{1}_{[S_j \neq 0]}$ runs separated with markers.

3) Fix $\varepsilon > 0$. If $K_n < k \triangleq n(\mathbb{E}[S] - \varepsilon)$ or $\hat{K}_n < \hat{k} \triangleq n(1-\delta-\varepsilon)$ declare error, whose probability is denoted by κ_n where k and $\hat{k}$ are assumed to be integers for computational simplicity. Otherwise, proceed with the next step.

4) Match the l^{th} row $Y_l^{K_n}$ of $\mathbf{D}^{(2)}$ X_i^n of $\mathbf{D}^{(1)}$, assigning $\hat{\Theta}_n(i) = l$, if i is the only index in $[m_n]$ such that *i)* X_i^n is ε-typical with respect to p_X and *ii)* $\tilde{X}_i^n$ contains a subsequence of length $\tilde{K}_n$, jointly ε-typical with $\tilde{Y}_l^K$ with respect to $p_{X,Y,\hat{S}}$ where $\hat{S} \sim p_{\hat{S}}$ with

$$p_{\hat{S}}(s) = \begin{cases} \frac{p_S(s)}{1-\delta} & \text{if } s \in \{1,\dots,s_{\max}\} \\ 0 & \text{otherwise} \end{cases} \tag{372}$$

and

$$\Pr(Y^S = y^S | X = x, \hat{S} = s) = \prod_{j=1}^{s} p_{Y|X}(y_j|x). \tag{373}$$

Otherwise, declare a *collision* error.

Since additional runs in $\mathbf{D}^{(2)}$ and additional columns in each run would decrease the collision probability, we have

$$\Pr(\text{collision between 1 and } i | K_n \geq k, \tilde{K}_n \geq \tilde{k}) \leq \Pr(\text{collision between 1 and } i || K_n = k, \tilde{K}_n = \tilde{k}) \tag{374}$$

for any $i \in [m_n] \setminus \{1\}$. Thus, for the sake of simplicity, we can focus on the case $K = k$ as it yields an upper bound on the error probability of our matching scheme.

Let $A_\varepsilon^{(n)}(X)$ denote the set of ε-typical (with respect to p_X) sequences of length n and $A_\varepsilon(X^{\hat{k}}|Y_l^k, \hat{S}^{\hat{k}})$ denote the set of sequences of length $\hat{k}$ jointly ε-typical (with respect to $p_{X,Y,\hat{S}}$) with Y_l^k conditioned on $\hat{S}^n$. For the matching rows X_1^n, Y_l^k of $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$, define the pairwise collision probability between X_1^n and X_i^n where $i \neq 1$ as

$$P_{\text{col,i}} \triangleq \Pr(X_i^n \in A_\varepsilon^{(n)}(X) \text{ and } \exists z^{\hat{k}} \in A_\varepsilon(X^{\hat{k}}|Y_l^k, \hat{S}^{\hat{k}}) \text{ which is a subsequence of } X_i^n.). \tag{375}$$

Therefore given the correct labeling for $Y_l^k \in \mathbf{D}^{(2)}$ is $X_1^n \in \mathbf{D}^{(1)}$, the probability of error P_e can be bounded as

$$\begin{aligned} P_e &\leq \Pr(\nexists z^{\hat{k}} : z^{\hat{k}} \in A_\varepsilon(X^{\hat{k}}|Y_l^k, \hat{S}^{\hat{k}}) \text{ and } z^{\hat{k}} \text{ is a subsequence of } X_1^n.) \\ &\quad + \Pr(X_1^n \notin A_\varepsilon^{(n)}(X)) + \sum_{i=2}^{2^{nR}} P_{\text{col,i}} + \kappa_n + \rho_n \qquad (376) \\ &\leq 2\varepsilon + \sum_{i=2}^{2^{nR}} P_{\text{col,i}} + \kappa_n + \rho_n \qquad (377) \\ &\leq 2\varepsilon + 2^{nR} P_{\text{col,2}} + \kappa_n + \rho_n \qquad (378) \end{aligned}$$

where (378) follows from the fact the the rows are *i.i.d.* and thus $P_{\text{col,i}} = P_{\text{col,2}}$, $\forall i \in [m_n] \setminus \{1\}$.

We now upper bound $P_{\text{col,2}}$. For any $z^{\hat{k}}$ define

$$T(z^{\hat{k}}) \triangleq \{x^n \in \mathfrak{X}^n : x^n \in A_\varepsilon^{(n)}(X),\ x^n \text{ contains } z^{\hat{k}}.\}. \tag{379}$$

Observe that for any $z^{\hat{k}} \in A_\varepsilon(X^{\hat{k}}|Y_l^k, \hat{S}^{\hat{k}})$, we have $z^{\hat{k}} \in A_\varepsilon^{(\hat{k})}(X)$. Furthermore, for a given deletion pattern with $n-\hat{k} = \Theta(n)$ deletions, WLOG $(\hat{k}+1,\ldots,n)$, the ε-typicality of $x^n = (x_1,\ldots,x_n)$ and $z^{\hat{k}} = (x_1,\ldots,x_{\hat{k}})$ with respect to p_X implies the $\tilde{\varepsilon}$-typicality of $(x_{\hat{k}+1},\ldots,x_n)$, where $\tilde{\varepsilon} = \frac{2-\delta-\varepsilon}{\delta+\varepsilon}\varepsilon$. Therefore for any $z^{\hat{k}} \in A_\varepsilon(X^{\hat{k}}|Y_l^k, \hat{S}^{\hat{k}})$, taking the union bound over all possible deletion patterns with $n-\hat{k}$ deletions, the cardinality of $T(z^{\hat{k}})$ can be upper bounded as

$$\begin{aligned}
|T(z^{\hat{k}})| &\leq \binom{n}{\hat{k}} |A_{\tilde{\varepsilon}}^{(n-\hat{k})}(X)| && (380)\\
&\leq 2^{nH_b(\frac{\hat{k}}{n})} |A_{\tilde{\varepsilon}}^{(n-\hat{k})}(X)| && (381)\\
&\leq 2^{nH_b(\frac{\hat{k}}{n})} 2^{(n-\hat{k})(H(X)+\tilde{\varepsilon})} && (382)\\
&= 2^{n\left[H_b(\frac{\hat{k}}{n})+(1-\frac{\hat{k}}{n})(H(X)+\tilde{\varepsilon})\right]} && (383)
\end{aligned}$$

Furthermore, for any $x^n \in T(z^{\hat{k}})$, since $T(z^{\hat{k}}) \subseteq A_\varepsilon^{(n)}(X)$, we have

$$p_{X^n}(x^n) \leq 2^{-n(H(X)-\varepsilon)} \tag{384}$$

and since the rows X_i^n of $\mathbf{D}^{(1)}$ are *i.i.d.*, we have

$$\Pr(X_2^n \in T(z^{\hat{k}}) | X_1^n \in T(z^{\hat{k}})) = \Pr(X_2^n \in T(z^{\hat{k}})). \tag{385}$$

Finally, we note that

$$|A_\varepsilon(X^{\hat{k}}|Y_l^k, \hat{S}^{\hat{k}})| \leq 2^{\hat{k}(H(X|Y^{\hat{S}},\hat{S})+\varepsilon)} \tag{386}$$

and

$$\begin{aligned}
H(X|Y^{\hat{S}},\hat{S}) &= \sum_{s=1}^{s_{\max}} p_{\hat{S}}(s) H(X|Y^{\hat{S}}, \hat{S}=s) && (387)\\
&= \frac{1}{1-\delta} \sum_{s=1}^{s_{\max}} p_S(s) H(X|Y^{\hat{S}}, \hat{S}=s) && (388)\\
&= \frac{1}{1-\delta} \left[\sum_{s=0}^{s_{\max}} p_S(s) H(X|Y^{\hat{S}}, \hat{S}=s) - \delta H(X|Y^S, S=0)\right] && (389)\\
&= \frac{1}{1-\delta} \left[H(X|Y^S,S) - \delta H(X)\right] && (390)\\
&= \frac{1}{1-\delta} \left[(1-\delta)H(X) - I(X;Y^S,S)\right] && (391)\\
&= H(X) - \frac{I(X;Y^S,S)}{1-\delta} && (392)
\end{aligned}$$

Thus, we get

$$|A_\varepsilon(X^{\hat{k}}|Y_l^k, \hat{S}^{\hat{k}})| \leq 2^{\hat{k}\left[H(X)-\frac{I(X;Y^S,S)}{1-\delta}+\varepsilon\right]}. \tag{393}$$

Combining (383)-(393), we can upper bound $P_{\text{col},2}$ as

$$P_{\text{col},2} \leq \sum_{z^{\hat{k}} \in A_\varepsilon(X^{\hat{k}}|Y_l^k,\hat{S}^{\hat{k}})} \Pr(X_2^n \in T(z^{\hat{k}})) \tag{394}$$

$$= \sum_{z^{\hat{k}} \in A_\varepsilon(X^{\hat{k}}|Y_l^k,\hat{S}^{\hat{k}})} \sum_{x^n \in T(z^{\hat{k}})} p_{X^n}(x^n) \tag{395}$$

$$\leq \sum_{z^{\hat{k}} \in A_\varepsilon(X^{\hat{k}}|Y_l^k,\hat{S}^{\hat{k}})} \sum_{x^n \in T(z^{\hat{k}})} 2^{-n(H(X)-\varepsilon)} \tag{396}$$

$$= \sum_{z^{\hat{k}} \in A_\varepsilon(X^{\hat{k}}|Y_l^k,\hat{S}^{\hat{k}})} |T(z^{\hat{k}})| 2^{-n(H(X)-\varepsilon)} \tag{397}$$

$$\leq \sum_{z^{\hat{k}} \in A_\varepsilon(X^{\hat{k}}|Y_l^k,\hat{S}^{\hat{k}})} 2^{-n(H(X)-\varepsilon)} 2^{n\left[H_b(\frac{\hat{k}}{n})+(1-\frac{\hat{k}}{n})(H(X)+\tilde{\varepsilon})\right]} \tag{398}$$

$$= |A_\varepsilon(X^{\hat{k}}|Y_l^k,\hat{S}^{\hat{k}})| 2^{-\left[\hat{k}H(X)-n\varepsilon-H_b(\frac{\hat{k}}{n})-(n-\hat{k})\tilde{\varepsilon}\right]} \tag{399}$$

$$\leq 2^{\hat{k}\left[H(X)-\frac{I(X;Y^S,S)}{1-\delta}+\varepsilon\right]} 2^{-\left[\hat{k}H(X)-n\varepsilon-nH_b(\frac{\hat{k}}{n})-(n-\hat{k})\tilde{\varepsilon}\right]} \tag{400}$$

$$= 2^{-n\left[\frac{1-\delta-\varepsilon}{1-\delta}I(X;Y^S,S)-H_b(\delta+\varepsilon)-(\delta+\varepsilon)(\varepsilon+\tilde{\varepsilon})\right]} \tag{401}$$

$$= 2^{-n\left[\frac{1-\delta-\varepsilon}{1-\delta}I(X;Y^S,S)-H_b(\delta+\varepsilon)-2\varepsilon\right]} \tag{402}$$

Thus, we have the following upper bound on the error probability

$$P_e \leq 2\varepsilon + 2^{nR} 2^{-n\left[\frac{1-\delta-\varepsilon}{1-\delta}I(X;Y^S,S)-H_b(\delta+\varepsilon)-2\varepsilon\right]} + \kappa_n + \rho_n \tag{403}$$

By LLN, we have $\kappa_n \to 0$ and from Lemma 1, we have $\rho_n \to 0$ as $n \to \infty$. Hence, we can argue that any database growth rate R satisfying

$$R < I(X;Y^S,S) - H_b(\delta) \tag{404}$$

is achievable by taking ε small enough.

Now, we investigate repetition distributions with $\delta \leq 1 - \frac{1}{|\mathfrak{X}|}$. Recall from Appendix D the counting function $F(n,\hat{k},|\mathfrak{X}|)$ denoting the number of $|\mathfrak{X}|$-ary sequences of length n, which contain a fixed $|\mathfrak{X}|$-ary sequence of length $\hat{k}$ as a subsequence. From [55], [63], we have

$$F(n,\hat{k},|\mathfrak{X}|) \leq n 2^{n\left[H_b\left(\frac{\hat{k}}{n}\right)+(1-\frac{\hat{k}}{n})\log(|\mathfrak{X}|-1))\right]}. \tag{405}$$

Furthermore, disregarding the typicality constraint, we can trivially bound the cardinality of $T(z^{\hat{k}})$ as

$$|T(z^{\hat{k}})| \leq |\{x^n \in \mathfrak{X}^n : x^n \text{ contains } z^{\hat{k}}\}| \tag{406}$$

$$\leq F(n,\hat{k},|\mathfrak{X}|) \tag{407}$$

$$\leq n 2^{n\left[H_b\left(\frac{\hat{k}}{n}\right)+(1-\frac{\hat{k}}{n}\log(|\mathfrak{X}|-1))\right]} \tag{408}$$

Plugging (408) into (397) and following the same steps, one can show that any rate R satisfying

$$R < \left[I(X;Y^S,S) + \delta(H(X) - \log(|\mathfrak{X}|-1)) - H_b(\delta)\right]^+ \tag{409}$$

is achievable. Simply taking the maximum of the two proven achievable rates when $\delta \leq 1 - 1/|\mathfrak{X}|$ yields the desired achievability result. This concludes the proof. □

APPENDIX I
PROOF OF LEMMA 4

For brevity, we let μ_n denote $\Pr(\exists i,j \in [n], i \neq j, H_i^{(1)} = H_j^{(1)})$. Notice that since the entries of $\mathbf{D}^{(1)}$ are *i.i.d.*, $H_i^{(1)}$ are *i.i.d.* Multinomial(m_n, p_X) random variables. Then,

$$\mu_n \leq n^2 \Pr(H_1^{(1)} = H_2^{(1)}) \tag{410}$$

$$= n^2 \sum_{h^{|\mathfrak{X}|}} \Pr(H_1^{(1)} = h^{|\mathfrak{X}|})^2 \tag{411}$$

where the sum is over all vectors of length $|\mathfrak{X}|$, summing up to m_n. Let $m_i \triangleq h(i)$, $\forall i \in \mathfrak{X}$. Then,

$$\Pr(H_1^{(1)} = h^{|\mathfrak{X}|}) = \binom{m_n}{m_1, m_2, \ldots, m_{|\mathfrak{X}|}} \prod_{i=1}^{|\mathfrak{X}|} p_X(i)^{m_i} \tag{412}$$

Hence, we have

$$\mu_n \leq n^2 \sum_{m_1+\cdots+m_{|\mathfrak{X}|}=m_n} \binom{m_n}{m_1, m_2, \ldots, m_{|\mathfrak{X}|}}^2 \prod_{i=1}^{|\mathfrak{X}|} p_X(i)^{2m_i} \tag{413}$$

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

From [49, Theorem 11.1.2], we have

$$\prod_{i=1}^{|\mathfrak{X}|} p_X(i)^{2m_i} = 2^{-2m_n(H(\tilde{p})+D(\tilde{p}\|p_X))} \tag{414}$$

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

$$\tilde{p} = \left(\frac{m_1}{m_n}, \ldots, \frac{m_{|\mathfrak{X}|}}{m_n}\right). \tag{415}$$

From Stirling's approximation [50, Chapter 3.2], we get

$$\binom{m_n}{m_1, m_2, \ldots, m_{|\mathfrak{X}|}}^2 \leq \frac{e^2}{(2\pi)^{|\mathfrak{X}|}} m_n^{1-|\mathfrak{X}|} \Pi_{\tilde{p}}^{-1} 2^{2m_n H(\tilde{p})} \tag{416}$$

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

Combining (413)-(416), we get

$$\mu_n \leq \frac{e^2}{(2\pi)^{|\mathfrak{X}|}} n^2 m_n^{1-|\mathfrak{X}|} \sum_{\tilde{p}} \Pi_{\tilde{p}}^{-1} 2^{-2m_n D_{KL}(\tilde{p}\|p_X)} \tag{417}$$

Let

$$\tilde{T} = \sum_{\tilde{p}} \Pi_{\tilde{p}}^{-1} 2^{-2m_n D_{KL}(\tilde{p}\|p_X)} = \tilde{T}_1 + \tilde{T}_2 \tag{418}$$

where

$$\tilde{T}_1 = \sum_{\tilde{p}: D_{KL}(\tilde{p}\|p_X) > \frac{\varepsilon_n^2}{2\log_e 2}} \Pi_{\tilde{p}}^{-1} 2^{-2m_n D_{KL}(\tilde{p}\|p_X)} \tag{419}$$

$$\tilde{T}_2 = \sum_{\tilde{p}: D_{KL}(\tilde{p}\|p_X) \leq \frac{\varepsilon_n^2}{2\log_e 2}} \Pi_{\tilde{p}}^{-1} 2^{-2m_n D_{KL}(\tilde{p}\|p_X)}, \tag{420}$$

ε_n, which is described below in more detail, is a small positive number decaying with n.

First, we look at $\tilde{T}_2$. From Pinsker's inequality [49, Lemma 11.6.1], we have

$$D_{KL}(\tilde{p}\|p_X) \leq \frac{\varepsilon_n^2}{2\log_e 2} \Rightarrow \mathrm{TV}(\tilde{p}, p_X) \leq \varepsilon_n \tag{421}$$

where TV denotes the total variation distance. Therefore

$$\begin{aligned}\left|\{\tilde{p} : D_{KL}(\tilde{p}\|p_X) \leq \frac{\varepsilon_n^2}{2\log_e 2}\}\right| &\leq |\{\tilde{p} : \mathrm{TV}(\tilde{p}, p_X) \leq \varepsilon_n\}| \\ &= O(m_n^{|\mathfrak{X}|-1}\varepsilon_n^{|\mathfrak{X}|-1})\end{aligned} \tag{422}$$

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

$$\Pi_{\tilde{p}} \geq \prod_{i=1}^{|\mathfrak{X}|}(p_X(i)-\varepsilon_n) \geq \Pi_{p_X} - \varepsilon_n \sum_{i=1}^{|\mathfrak{X}|}\prod_{j\neq i} p_X(j) \tag{423}$$

Hence

$$\Pi_{\tilde{p}}^{-1} \leq \frac{1}{\Pi_{p_X} - \varepsilon_n \sum\limits_{i=1}^{|\mathfrak{X}|}\prod\limits_{j\neq i} p_X(j)} \tag{424}$$

and

$$\tilde{T}_2 \leq \frac{1}{\Pi_{p_X} - \varepsilon_n \sum\limits_{i=1}^{|\mathfrak{X}|}\prod\limits_{j\neq i} p_X(j)} O(m_n^{|\mathfrak{X}|-1}\varepsilon_n^{|\mathfrak{X}|-1}) \tag{425}$$

$$= O(m_n^{|\mathfrak{X}|-1}\varepsilon_n^{|\mathfrak{X}|-1}) \tag{426}$$

for small ε_n.

Now, we look at $\tilde{T}_1$. Note that since $m_i \in \mathbb{Z}_+$, we have $\Pi_{\tilde{p}} \leq m_n^{|\mathfrak{X}|}$, suggesting the multiplicative term in the summation in (419) is polynomial with m_n. If $m_i = 0$ we can simply discard it and return to Stirling's approximation with the reduced number of categories. Furthermore, from [49, Theorem 11.1.1], we have

$$\left|\{\tilde{p} : D_{KL}(\tilde{p}\|p_X) > \frac{\varepsilon_n^2}{2\log_e 2}\}\right| \leq |\{\tilde{p}\}| \tag{427}$$

$$\leq (m_n+1)^{|\mathfrak{X}|} \tag{428}$$

suggesting the number of terms which we take the summation over in (419) is polynomial with m_n as well. Therefore, as long as $m_n\varepsilon_n^2 \to \infty$, $\tilde{T}_1$ has a polynomial number of elements which decay exponentially with m_n. Thus

$$\tilde{T}_1 \to 0 \text{ as } n \to \infty. \tag{429}$$

Define

$$U_i = e^2(2\pi)^{-|\mathfrak{X}|}m_n^{1-|\mathfrak{X}|}\tilde{T}_i, \quad i = 1,2 \tag{430}$$

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

$$U_2 = O(\varepsilon_n^{|\mathfrak{X}|-1}) = O(m_n^{(1-|\mathfrak{X}|)/2}V_n^{(|\mathfrak{X}|-1)}). \tag{431}$$

Combining (429)-(431), we have

$$U = U_1 + U_2 = O(m_n^{(1-|\mathfrak{X}|)/2} V_n^{(|\mathfrak{X}|-1)}) \tag{432}$$

and we get

$$\mu_n \leq n^2 O(m_n^{(1-|\mathfrak{X}|)/2} V_n^{(|\mathfrak{X}|-1)}) \tag{433}$$

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

$$\mu_n \leq O(n^2 n^{-2} Z_n^{(1-|\mathfrak{X}|)/2} V_n^{(|\mathfrak{X}|-1)}) = o(1) \tag{434}$$

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

APPENDIX J
PROOF OF PROPOSITION 1

For brevity, we let μ_n denote $\Pr(\exists i, j \in [n], i \neq j, H_i^{(1)} = H_j^{(1)})$. Then,

$$\begin{aligned}
\mu_n &= n(n-1) \Pr(H_1^{(1)} = H_2^{(1)}) && (435)\\
&= n(n-1) \sum_{h^{|\mathfrak{X}|}} \Pr(H_1^{(1)} = h^{|\mathfrak{X}|})^2 && (436)\\
&= n(n-1) \sum_{m_1+\cdots+m_{|\mathfrak{X}|}=m_n} \binom{m_n}{m_1, \ldots, m_{|\mathfrak{X}|}}^2 |\mathfrak{X}|^{-2m_n} && (437)\\
&= n(n-1)|\mathfrak{X}|^{-2m_n} \sum_{m_1+\cdots+m_{|\mathfrak{X}|}=m_n} \binom{m_n}{m_1, \ldots, m_{|\mathfrak{X}|}}^2 && (438)\\
&= n(n-1)|\mathfrak{X}|^{|\mathfrak{X}|/2} (4\pi m_n)^{(1-|\mathfrak{X}|)/2} (1+o_{m_n}(1))(1-o_n(1)) && (439)\\
&= n^2 m_n^{\frac{1-|\mathfrak{X}|}{2}} (4\pi)^{(1-|\mathfrak{X}|)/2} |\mathfrak{X}|^{|\mathfrak{X}|/2} (1+o_{m_n}(1))(1-o_n(1)) && (440)
\end{aligned}$$

where (439) follows from [65, Theorem 4]. □