Matching of Markov Databases Under Random Column Repetitions

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

***Abstract*—Matching entries of correlated shuffled databases have practical applications ranging from privacy to biology. In this paper, motivated by synchronization errors in the sampling of time-indexed databases, matching of random databases under random column repetitions and deletions is investigated. It is assumed that for each entry (row) in the database, the attributes (columns) are correlated, which is modeled as a Markov process. Column histograms are proposed as a permutation-invariant feature to detect the repetition pattern, whose asymptotic-uniqueness is proved using information-theoretic tools. Repetition detection is then followed by a typicality-based row matching scheme. Considering this overall scheme, sufficient conditions for successful matching of databases in terms of the database growth rate are derived. A modified version of Fano's inequality leads to a tight necessary condition for successful matching, establishing the matching capacity under column repetitions. This capacity is equal to the erasure bound, which assumes the repetition locations are known a-priori. Overall, our results provide insights on privacy-preserving publication of anonymized time-indexed data.**

I. Introduction

Recently, with the proliferation of smart devices and the emergence of big data applications, 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 literatures through successful practical de-anonymization attacks on real data [3]–[16].

In the light of the above practical attacks, several groups initiated rigorous analyses of the graph matching problem [17]–[26]. Correlated graph matching has applications beyond privacy, such as image processing [27] and DNA sequencing, which is shown to be equivalent to matching bipartite graphs [28]. Matching of correlated databases, also equivalent to bipartite graph matching, have been investigated from information-theoretic [29]–[33] and statistical [34] perspectives. In [30], Cullina *et al.* introduced *cycle mutual information* as a correlation metric and derived sufficient conditions for successful matching and a converse result using perfect recovery as error criterion. In [29], Shirani *et al.* considered a pair of databases of the same size, and drawing an analogy between channel decoding and database matching, derived necessary and sufficient conditions on the database growth rate for successful database matching. In [31], Dai *et al.* considered the matching of a pair of databases with jointly Gaussian features with perfect recovery constraint. Similarly, in [34], Kunisky and Niles-Weed considered the same problem from the statistical perspective in different regimes of database size and under several recovery criteria.

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

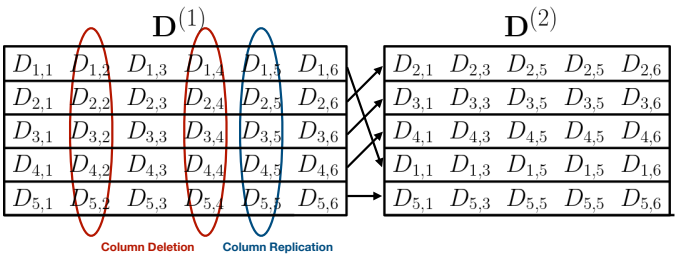

Fig. 1. An illustrative example of database matching under column repetitions. The columns circled in red are deleted whereas the column circled in blue is repeated twice, *i.e.,* replicated. Our goal is to estimate the row permutation Θ, which in this example given as; $\Theta(1)=4$, $\Theta(2)=1$,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

Motivated by synchronization errors in sampling of time-series datasets, in our prior work we considered database matching under *random column deletions* [32]. Assuming an *i.i.d.* underlying distribution for database attributes (columns) and the same successful matching criterion as [29], we derived an achievable database growth rate assuming a probabilistic side information on the deletion locations. We also proposed an algorithm to extract the side information on the deletion locations from a batch of already-matched rows, called *seeds.*

In this paper, we generalize [32], by assuming a more general model for synchronization errors, namely *column repetitions* where in addition to some columns being deleted as in [32], some columns may be sampled several times consecutively, *i.e., replicated*, as illustrated in Figure 1. Furthermore, in order to account for the potential correlation among the attributes (columns), we model the rows using a Markov process contrary. Under this generalized model, we derive an improved achievable database growth rate. We propose a novel *histogram-based repetition detection* algorithm, where we compare the column histograms in order to infer the column repetition pattern with high probability, followed by a typicality-based matching scheme to match the rows of the correlated databases. We also derive the necessary conditions for successful matching. We show that the necessary and sufficient conditions are tight up to equality, and equal to

the erasure bound, which is obtained when there are no replications and the deletion locations are perfectly known. Thus, we completely characterize the *capacity* of the matching of column repeated databases.

The organization of this paper is as follows: Section II introduces the problem formulation. In Section III, our main result on matching capacity and the proof of the achievability are presented. In Section IV, the converse is proved. Finally, in Section V the results and ongoing work are discussed.

Notation: We denote the set of integers $\{1,...,n\}$ as $[n]$, and matrices with uppercase bold letters. For a matrix $\mathbf{D}$, $D_{i,j}$ denotes the $(i,j)^{\text{th}}$ entry. Furthermore, by A^n, we denote a row vector consisting of scalars $A_1,\ldots,A_n$ and the indicator of event ε by $\mathbb{1}_\varepsilon$. The logarithms, unless stated explicitly, are in base 2.

II. Problem Formulation

We use the following definitions, some of which are similar to [29], [32], [33] to formalize 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 $\mathbf{D} = \{D_{i,j} \in \mathfrak{X} : i \in [m_n], j \in [n]\}$ 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

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

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

and

$$\sum_{j\in\mathfrak{X}} u_j = 1 \tag{3}$$

$$\gamma \in \Big(-\min_{j\in\mathfrak{X}} \frac{u_j}{1-u_j}, 1\Big) \tag{4}$$

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

Definition 2. (Column Repetition Pattern) The *column repetition pattern* S^n is a random vector consisting of *i.i.d.* elements S_j, $j \in [n]$, drawn from a discrete probability distribution p_S with a finite discrete support $\{0,\ldots,s_{\max}\}$. The parameter $\delta \triangleq p_S(0)$ is called the *deletion probability*.

Definition 3. (Labeled Repeated Database) Let $\mathbf{D}^{(1)}$ be an (m_n,n,P) unlabeled Markov database, S^n be the column repetition pattern, and Θ_n be a uniform permutation of $[m_n]$ with $\mathbf{D}^{(1)}$, S^n and Θ_n independently chosen. Given $\mathbf{D}^{(1)}$ and S^n, the pair $(\mathbf{D}^{(2)},\Theta_n)$ is called the *labeled repeated database* if the $(i,j)^{\text{th}}$ element $D^{(1)}_{i,j}$ of $\mathbf{D}^{(1)}$ and its counterpart $D^{(2)}_{i,j}$ in $\mathbf{D}^{(2)}$ have the following relation:

$$D^{(2)}_{i,j} = \begin{cases} E, & \text{if } S_j = 0 \\ D^{(1)}_{\Theta_n^{-1}(i),j} \otimes 1^{S_j} & \text{if } S_i \geq 1 \end{cases} \tag{5}$$

where 1^{S_j} and $\otimes$ denote the all-ones row vector of length S_j and the Kronecker product, respectively. Furthermore $D^{(2)}_{i,j} = E$ corresponds to the empty string and $D^{(2)}_{i,j} = D^{(1)}_{\Theta_n^{-1}(i),j} \otimes 1^{S_j}$ corresponds to the j^{th} column of $\mathbf{D}^{(2)}$ being an $m_n \times S_j$ matrix consisting of S_j copies of the j^{th} column of $\mathbf{D}^{(1)}$, concatenated together after shuffling with Θ_n. Θ_n and $\mathbf{D}^{(2)}$ are called the *labeling function* and *correlated column repeated database*, respectively. The respective rows $D^{(1)}_{i_1}$ and $D^{(2)}_{i_2}$ of $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$ are said to be *matching rows*, if $\Theta_n(i_1) = i_2$.

If $S_j = 0$, the j^{th} column of $\mathbf{D}^{(1)}$ is said to be *deleted* and if $S_j > 1$, j^{th} column of $\mathbf{D}^{(1)}$ is said to be *replicated*.

In our model, the correlated column 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 column repetition based on the repetition pattern S^n. We further assume that there is no noise on the retained entries, as is often done in the repeat channel literature [35].

Definition 4. (Successful Matching Scheme) A *matching scheme* is a sequence of mappings $\varphi_n : (\mathbf{D}^{(1)},\mathbf{D}^{(2)}) \to \hat{\Theta}_n$ where $\mathbf{D}^{(1)}$ is the unlabeled database, $\mathbf{D}^{(2)}$ is the correlated column repeated database and $\hat{\Theta}_n$ is the estimate of the true labeling function Θ_n. The scheme φ_n is said to be *successful* if

$$\lim_{n\to\infty} \Pr\big(\Theta_n(J) \neq \hat{\Theta}_n(J)\big) \to 0 \tag{6}$$

where $J \sim \text{Unif}([m_n])$. Here the event $\Theta_n(J) \neq \hat{\Theta}_n(J)$ is called the *matching error*.

Similar to [29], [32], [33], our performance metric is the probability of mismatch of a uniformly chosen row. This formulation allows us to derive results for a wide set of database distributions. Note that this performance metric is different than those of [30], [31], where the probability of the perfect recovery of the complete labeling function Θ_n was considered.

The relationship between the row size m_n and the column size n of the unlabeled database affects the probability of matching error in the following fashion: For a given n, as m_n increases, so does the probability of matching error due to the increased number of candidate rows. As stated in [34, Theorem 1.2], for the setting in our paper, the regime of interest is m_n growing exponentially in n.

Definition 5. (Database Growth Rate) The *database growth rate* R of an unlabeled Markov database with m_n rows and n columns is defined as

$$R = \lim_{n\to\infty} \frac{1}{n} \log m_n \tag{7}$$

Definition 6. (Achievable Database Growth Rate) Consider a sequence of (m_n,n,P) unlabeled Markov databases, a repetition probability distribution p_S and the resulting labeled repeated databases. 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 7. (Matching Capacity) The *matching capacity* C is the supremum of the set of all achievable rates corresponding to a probability transition matrix $\mathbf{P}$ and a repetition probability distribution p_S.

Our goal in this paper is to characterize the matching capacity of the database matching problem under the aforementioned Markovian row process and random column repetition models.

III. Matching Capacity and Achievability

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

Theorem 1. ***(Matching Capacity Under Column Repetitions)*** *Consider a probability transition matrix* $\mathbf{P}$ *and a repetition probability distribution* p_S*. Then, the matching capacity is*

$$C = (1-\delta)^2 \sum_{r=0}^{\infty} \delta^r H(X_0|X_{-r-1}) \tag{8}$$

where $\delta \triangleq p_S(0)$ *is the deletion probability and* $H(X_0|X_{-r-1})$ *is the entropy rate associated with the probability transition matrix*

$$\mathbf{P}^{r+1} = \gamma^{r+1}\mathbf{I} + (1-\gamma^{r+1})\mathbf{U}/ \tag{9}$$

The capacity can further be simplified as

$$\begin{aligned} C = \frac{(1-\delta)(1-\gamma)}{(1-\gamma\delta)}[H(\pi) + \sum_{i\in\mathfrak{X}} u_i^2 \log u_i] \\ - (1-\delta)^2 \sum_{r=0}^{\infty} \delta^r \sum_{i\in\mathfrak{X}} u_i(\gamma^{r+1} + (1-\gamma^{r+1})u_i) \\ \log(\gamma^{r+1} + (1-\gamma^{r+1})u_i) \end{aligned} \tag{10}$$

where $H(\pi)$ *denotes the entropy of the stationary distribution* π*.*

Observe that the RHS of (8) is the mutual information rate for an erasure channel with erasure probability δ with first-order Markov ($\mathbf{P}$) inputs, as given in [36]. Thus, Theorem 1 states that we can achieve the erasure bound which assumes a-priori knowledge of the column repetition pattern. This is unlike channel synchronization problems, where the erasure bound is a loose upper bound on the capacity. As we see below, the identicality of the repetition pattern across rows allows us to detect repetitions using a collapsed histogram based detection applied to columns. Finally, the matching capacity depends on the repetition distribution only through the deletion probability $\delta = p_S(0)$, rendering the replicated columns of the database irrelevant, as discussed in Section IV.

Note that the special case where $\gamma = 0$ results in an *i.i.d.* database distribution. Thus we have the following corollary:

Corollary 1. ***(i.i.d. Database Columns)*** *When the database entries* $D_{i,j}$ *are drawn* i.i.d. *from* $\mathfrak{X}$ *according to* p_X*, the matching capacity becomes*

$$C = (1-\delta)H(X) \tag{11}$$

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

Note that Corollary 1 improves the achievability result of [32]. Therefore, in addition to generalizing [32] to Markov databases and column repetitions, this paper also improves the achievability result of [32].

To prove the achievability of the matching capacity in Theorem 1, we consider the following two-phase matching scheme: Given $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$, the unlabeled and the correlated column repeated databases, we first infer the underlying repetition pattern S^n using the *collapsed histogram based detection* algorithm on the column histograms of $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$. Then, we use a joint typicality based sequence matching scheme to match the rows of $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$.

The collapsed histogram based 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:

$$\tilde{\mathbf{D}}_{i,j}^{(r)} = \begin{cases} 1 & \text{if } \mathbf{D}_{i,j}^{(r)} = 1 \\ 2 & \text{if } \mathbf{D}_{i,j}^{(r)} \neq 1 \end{cases}, \forall (i,j),\, r = 1,2 \tag{12}$$

From [37, Theorem 3] and (1), 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:

$$\tilde{\mathbf{P}} = \begin{bmatrix} \gamma + (1-\gamma)u_1 & (1-\gamma)(1-u_1) \\ (1-\gamma)u_1 & 1-(1-\gamma)u_1 \end{bmatrix} \tag{13}$$

$$\tilde{\pi} = \begin{bmatrix} u_1 & 1-u_1 \end{bmatrix} \tag{14}$$

Note that after collapsing the Markov chain, the histogram of the j^{th} column of $\tilde{\mathbf{D}}^{(1)}$ can be represented by the scalar $\tilde{H}_j^{(1)}$ which denotes the number of occurrences of state 2 in the j^{th} column of $\tilde{\mathbf{D}}^{(1)}$, $j \in [n]$. More formally, we have

$$\tilde{H}_j^{(1)} \triangleq \sum_{i=1}^{m_n} \mathbb{1}_{\left[\tilde{D}_{i,j}^{(1)}=2\right]}, \forall j \in [n] \tag{15}$$

Our histogram-based detection algorithm exploits two facts: First, the histogram (equivalently the type) of each column of $\tilde{\mathbf{D}}^{(1)}$ and $\tilde{\mathbf{D}}^{(2)}$ is invariant to row permutations. Second, as we prove in Lemma 1, the histogram of each column is asymptotically-unique due to the row size m_n being exponential in the column size n. Finally, since there is no noise on the retained entries of $\mathbf{D}^{(2)}$, we can match the column histograms, present in both $\tilde{\mathbf{D}}^{(1)}$ and $\tilde{\mathbf{D}}^{(2)}$ and detect the deleted columns, in an error-free fashion.

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

Lemma 1. ***(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* (15)*. Then,*

$$\Pr\left(\exists i,j \in [n],\, i \neq j, \tilde{H}_i^{(1)} = \tilde{H}_j^{(1)}\right) \to 0 \text{ as } n \to \infty \tag{16}$$

if $m_n = \omega(n^4)$.

Proof. See Appendix A. □

Note that by Definition 5, m_n is exponential in n and the order relation of Lemma 1 is automatically satisfied.

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

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

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

$$\tilde{H}_j^{(r)} = \sum_{i=1}^{m_n} \mathbb{1}_{\left[\tilde{D}_{i,j}^{(r)}=2\right]}, \quad \begin{cases} \forall j \in [n], & \text{if } r=1 \\ \forall j \in [K] & \text{if } r=2 \end{cases} \tag{17}$$

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 $\tilde{H}^{(2),K}$, 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 $\tilde{H}^{(2),K}$, 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 $p_{Y|X}$ be the probability transition matrix of an erasure channel with erasure probability δ, that is

$$p_{Y|X}(y|x) = \begin{cases} 1-\delta & \text{if } y = x \\ \delta & \text{if } y = \varepsilon \end{cases}, \quad \forall (x,y) \in \mathfrak{X}^2 \tag{18}$$

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 [38, Chapter 3] with $\bar{Y}_l^n$ with respect to p_{X^n,Y^n}, where

$$p_{X^n,Y^n}(x^n,y^n) = p_{X^n}(x^n) \prod_{j=1}^{n} p_{Y|X}(y_j|x_j) \tag{19}$$

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

Denote the ε-typical set of sequences (with respect to p_{X^n}) by $A_\varepsilon^{(n)}(X)$ and the jointly ε-typical set of sequences (with respect to p_{X^n,Y^n}) by $A_\varepsilon^{(n)}(X,Y)$. Supposing the true label for the l^{th} row $\bar{Y}_l^n$ of $\bar{\mathbf{D}}^{(2)}$ is 1, *i.e.,* $\Theta_n(1) = l$, and denoting the pairwise collision probability between X_1^n and X_i^n, by $P_{col,i}$, for any $i \neq 1$ we have

$$P_{col,i} = \Pr((X_i^n, \bar{Y}_l^n) \in A_\varepsilon^{(n)}) \tag{20}$$

$$\leq 2^{-n(I(X;Y)-3\varepsilon)} \tag{21}$$

where

$$I(X;Y) = \lim_{n\to\infty} \frac{I(X^n;\bar{Y}^n)}{n} \tag{22}$$

is the mutual information rate of the joint probability distribution p_{X^n,Y^n}. Thus, we can bound the probability of error P_e as

$$P_e \leq \mu_n + \Pr(X_1^n \notin A_\varepsilon^{(n)}(X)) + \sum_{i=2}^{n} P_{col,i} \tag{23}$$

$$\leq \mu_n + \varepsilon + \sum_{i=2}^{n} 2^{-n(I(X;Y)-3\varepsilon)} \tag{24}$$

$$= \mu_n + \varepsilon + 2^{n(R-I(X;Y)+3\varepsilon)} \tag{25}$$

Since m_n is exponential in n, by Lemma 1, $\mu_n \to 0$ as $n \to \infty$. Thus

$$P_e < 3\varepsilon \text{ as } n \to \infty \tag{26}$$

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

$$R < I(X;Y) \tag{27}$$

is achievable, by taking ε small enough. From [36, Corollary II.2] we have

$$I(X;Y) = (1-\delta)^2 \sum_{r=0}^{\infty} \delta^r H(X_0|X_{-r-1}) \tag{28}$$

where $H(X_0|X_{-r-1})$ is the entropy rate associated with the probability transition matrix $\mathbf{P}^{r+1}$. Finally, we prove (9) through induction. By Definition 1, (9) is satisfied for $r = 0$. Now assume (9) is true for some $r \in \mathbb{N}$. In other words,

$$\mathbf{P}^r = \gamma^r \mathbf{I} + (1-\gamma^r)\mathbf{U} \tag{29}$$

Observing $\mathbf{U}^k = \mathbf{U}$, $\forall k \in \mathbb{N}$, we obtain

$$\mathbf{P}^{r+1} = (\gamma\mathbf{I} + (1-\gamma)\mathbf{U})(\gamma^r\mathbf{I} + (1-\gamma^r)\mathbf{U}) \tag{30}$$

$$= \gamma^{r+1}\mathbf{I} + (\gamma(1-\gamma^r) + (1-\gamma)\gamma^r)\mathbf{U} + (1-\gamma)(1-\gamma^r)\mathbf{U}^2 \tag{31}$$

$$= \gamma^{r+1}\mathbf{I} + (\gamma(1-\gamma^r) + (1-\gamma)\gamma^r + (1-\gamma)(1-\gamma^r))\mathbf{U} \tag{32}$$

$$= \gamma^{r+1}\mathbf{I} + (1-\gamma^{r+1})\mathbf{U} \tag{33}$$

From (28)-(33) and [38, Theorem 4.2.4] we obtain (10), concluding the achievability part of the proof. □

IV. Converse

Theorem 1 states that we can convert repetitions to erasures, achieving the erasure bound. In this section, we show that the lower bound on the matching capacity C given in Section III is in fact tight, by proving it to also be an upper bound on the matching capacity C.

Proof of Converse of Theorem 1. Here we prove that the erasure bound given in (8) is an upper bound on all achievable database growth rates. We adopt a genie-aided proof where the repetition pattern S^n is available a-priori. Furthermore, we use the modified Fano's inequality presented in [29].

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,

$$P_e \triangleq \Pr\left(\Theta_n(J) \neq \hat{\Theta}_n(J)\right), \quad J \sim \text{Unif}([m_n]) \tag{34}$$

Furthermore, let S^n be the repetition pattern and $K = \sum_{j=1}^{n} S_j$.

Since Θ_n is a uniform permutation, from Fano's inequality, we have

$$H(\Theta_n|\mathbf{D}^{(1)},\mathbf{D}^{(2)}) \leq 1 + P_e \log(m_n!) \tag{35}$$
$$\leq 1 + P_e m_n \log m_n \tag{36}$$

where (36) follows from $m_n! \leq m_n^{m_n}$. Thus, we get

$$H(\Theta_n) = H(\Theta_n|\mathbf{D}^{(1)},\mathbf{D}^{(2)}) + I(\Theta_n;\mathbf{D}^{(1)},\mathbf{D}^{(2)}) \tag{37}$$
$$\leq 1 + P_e m_n \log m_n + I(\Theta_n;\mathbf{D}^{(1)},\mathbf{D}^{(2)}) \tag{38}$$

Note that

$$I(\Theta_n;\mathbf{D}^{(1)},\mathbf{D}^{(2)}) = I(\Theta_n;\mathbf{D}^{(2)}) + I(\Theta_n;\mathbf{D}^{(1)}|\mathbf{D}^{(2)}) \tag{39}$$
$$= I(\Theta_n;\mathbf{D}^{(1)}|\mathbf{D}^{(2)}) \tag{40}$$
$$\leq I(\Theta_n,\mathbf{D}^{(2)};\mathbf{D}^{(1)}) \tag{41}$$

where in (40) we have used the independence of Θ_n and $\mathbf{D}^{(2)}$.

$$I(\Theta_n,\mathbf{D}^{(2)};\mathbf{D}^{(1)}) \leq I(\Theta_n,\mathbf{D}^{(2)},\mathbf{S}^n;\mathbf{D}^{(1)}) \tag{42}$$
$$= I(\mathbf{D}^{(1)};\Theta_n,\mathbf{D}^{(2)}|S^n) \tag{43}$$
$$= \sum_{i=1}^{m_n} I(D_i^{(1),n};D^{(2),K}_{\Theta_n^{-1}(i)}|S^n) \tag{44}$$
$$= m_n I(D_1^{(1),n};D^{(2),K}_{\Theta_n^{-1}(1)}|S^n) \tag{45}$$
$$= m_n I(D_1^{(1),n};D^{(2),K}_{\Theta_n^{-1}(1)},S^n) \tag{46}$$

where (43)-(46) follow from the fact that $\mathbf{D}^{(1)}$ and S^n are independent and non-matching rows are *i.i.d.* conditioned on the repetition pattern S^n.

Now, for brevity let $X^n = D_1^{(1),n}$, $Y^K = D^{(2),K}_{\Theta_n^{-1}(1)}$ and $\tilde{Y}^n$ be obtained from Y^K as described in Step 5 of the achievability proof. Since there is a bijective mapping between (Y^K,S^n) and $(\bar{Y}^n,S^n)$, we have

$$I(X^n;Y^K,S^n) = I(X^n;\bar{Y}^n,S^n) \tag{47}$$
$$= I(X^n;\bar{Y}^n) + I(X^n;S^n|\bar{Y}^n) \tag{48}$$
$$= I(X^n;\bar{Y}^n) \tag{49}$$

where (49) follows from the independence of S^n from 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.

Finally, from Stirling's approximation and the uniformity of Θ_n, we have

$$\lim_{n\to\infty} \frac{1}{m_n n} H(\Theta_n) = \lim_{n\to\infty} \left[\frac{1}{n}\log m_n + \frac{1}{m_n n} O(n)\right] = R \tag{50}$$

Therefore, from (38)-(50), we have

$$\lim_{n\to\infty} \frac{1}{m_n n} H(\Theta_n) \leq \lim_{n\to\infty} \left[\frac{1}{m_n n} + P_e \frac{1}{n}\log m_n + I(X^n;\bar{Y}^n)\right] \tag{51}$$
$$R \leq \lim_{n\to\infty} \frac{I(X^n;\bar{Y}^n)}{n} \tag{52}$$
$$= (1-\delta)^2 \sum_{r=0}^{\infty} \delta^r H(X_0|X_{-r-1}) \tag{53}$$

where (52) follows from the fact that $P_e \to 0$ as $n \to \infty$ and (53) follows from [36, Corollary II.2]. The rest of the proof follows from the evaluation of (53) we did in Section III. □

Equations (47)-(49) suggest that the additional copies of the replicated columns do not offer any information. As a result, discarding the additional replicas in the matching scheme of Section III does not impact optimality.

V. Conclusion

In this paper, we have studied the matching of Markov databases under random column repetitions. By proving the asymptotic-uniqueness of the column histograms of the databases, we have showed that these histograms can be used for the detection of the deleted and replicated columns. Using the proposed histogram-based detection and typicality-based row matching, we have derived an achievability result for database growth rate, which we have showed is tight, thus giving us the database matching capacity. Our ongoing work includes investigating the matching capacity in the presence of noise as well as synchronization errors [33] and when different subsets of rows are sampled separately and then merged together, *i.e.,* different subsets of rows experience different repetition patterns.

References

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

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

[3] F. M. Naini, J. Unnikrishnan, P. Thiran, and M. Vetterli, "Where you are is who you are: User identification by matching statistics," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 2, pp. 358–372, 2016.

[4] A. Datta, D. Sharma, and A. Sinha, "Provable de-anonymization of large datasets with sparse dimensions," in *International Conference on Principles of Security and Trust*. Springer, 2012, pp. 229–248.

[5] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. of IEEE Symposium on Security and Privacy*, 2008, pp. 111–125.

[6] L. Sweeney, "Weaving technology and policy together to maintain confidentiality," *The Journal of Law, Medicine & Ethics*, vol. 25, no. 2-3, pp. 98–110, 1997.

[7] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Matching anonymized and obfuscated time series to users' profiles," *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 724–741, 2019.
[8] 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, "Perfwed: 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] 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.
[18] L. Babai, P. Erdos, and S. M. Selkow, "Random graph isomorphism," *SIAM Journal on computing*, vol. 9, no. 3, pp. 628–635, 1980.
[19] S. Janson, A. Rucinski, and T. Luczak, *Random graphs*. John Wiley & Sons, 2011.
[20] T. Czajka and G. Pandurangan, "Improved random graph isomorphism," *Journal of Discrete Algorithms*, vol. 6, no. 1, pp. 85–92, 2008.
[21] 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.
[22] 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.
[23] 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.
[24] 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.
[25] 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.
[26] 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.
[27] 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.
[28] 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.
[29] F. Shirani, S. Garg, and E. Erkip, "A concentration of measure approach to database de-anonymization," in *Proc. of IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 2748–2752.
[30] 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.
[31] 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.
[32] S. Bakırtaş and E. Erkip, "Database matching under column deletions," in *Proc. of IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 2720–2725.
[33] S. Bakirtas and E. Erkip, "Seeded database matching under noisy column repetitions," *arXiv preprint arXiv:2202.01724*, 2022.
[34] D. Kunisky and J. Niles-Weed, "Strong recovery of geometric planted matchings," in *Proc. of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 2022, pp. 834–876.
[35] 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.
[36] Y. Li and G. Han, "Input-constrained erasure channels: Mutual information and capacity," in *Proc. of IEEE International Symposium on Information Theory (ISIT)*, 2014, pp. 3072–3076.
[37] 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.
[38] T. M. Cover, *Elements of Information Theory*. John Wiley & Sons, 2006.

APPENDIX

A. Proof of Lemma 1

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

$$\mu_n \leq \sum_{(i,j)\in[n]^2:i<j} \Pr(\tilde{H}_i^{(1)} = \tilde{H}_j^{(1)}) \tag{54}$$

$$\leq n^2 \max_{(i,j)\in[n]^2:i<j} \Pr(\tilde{H}_i^{(1)} = \tilde{H}_j^{(1)}) \tag{55}$$

Due to stationarity, the 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

$$q = \sum_{r=0}^{m_n} \Pr(\tilde{H}_1^{(1)} = r) \Pr(\tilde{H}_{s+1}^{(1)} = r | \tilde{H}_1^{(1)} = r) \tag{56}$$

$$= \sum_{r=0}^{m_n} \binom{m}{r} (1-u_1)^r u_1^{m_n - r} \Pr(\tilde{H}_{s+1}^{(1)} = r | \tilde{H}_1^{(1)} = r) \tag{57}$$

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

$$\Pr(\tilde{H}_{s+1}^{(1)} = r | \tilde{H}_1^{(1)} = r) = \Pr(A + B = r) \tag{58}$$

where $A \sim \mathrm{Binom}(r, Q_{1,1})$ and $B \sim \mathrm{Binom}(m_n - r, Q_{2,1})$ are independent. Then, from Stirling's approximation and [38, Theorem 11.1.2], we get

$$q = \sum_{r=0}^{m_n} \binom{m_n}{r} (1-u_1)^r u_1^{m_n - r} \Pr(A + B = r) \tag{59}$$

$$\leq \frac{e}{\sqrt{2\pi}} {m_n}^{-1/2} \sum_{r=0}^{m_n} \Pi_r^{-1} 2^{-m_n D(\frac{r}{m_n} \| (1-u_1))} \Pr(A + B = r) \tag{60}$$

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

$$T = \sum_{r=0}^{m_n} \Pi_r^{-1} 2^{-m_n D(\frac{r}{m_n}\|(1-u_1))} \Pr(A+B=r) = T_1 + T_2 \quad (61)$$

where

$$T_1 = \sum_{r: D(\frac{r}{m_n}\|1-u_1) > \frac{\varepsilon_n^2}{2\log_e 2}} \Pi_r^{-1} 2^{-m_n D(\frac{r}{m_n}\|(1-u_1))} \Pr(A+B=r) \quad (62)$$

$$T_2 = \sum_{r: D(\frac{r}{m_n}\|1-u_1) \le \frac{\varepsilon_n^2}{2\log_e 2}} \Pi_r^{-1} 2^{-m_n D(\frac{r}{m_n}\|(1-u_1))} \Pr(A+B=r), \quad (63)$$

$D(\frac{r}{m_n}\|(1-u_1))$ denotes the Kullback-Leibler divergence between Bernoulli($\frac{r}{m_n}$) and Bernoulli($1-u_1$) distributions, and $\varepsilon_n > 0$, 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 $\Pi_r \le {m_n}^2$, suggesting the multiplicative term in the summation in (62) 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

$$T_1 \to 0 \text{ as } n \to \infty \quad (64)$$

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

Now, we focus on T_2. From Pinsker's inequality [38, Lemma 11.6.1], we have

$$D\left(\frac{r}{m_n}\middle\|1-u_1\right) \le \frac{\varepsilon_n^2}{2\log_e 2} \Rightarrow \mathrm{TV}\left(\frac{r}{m_n}, 1-u_1\right) \le \varepsilon_n \quad (65)$$

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

$$\left|\{r : D\Big(\frac{r}{m_n}\Big\|1-u_1\Big) \le \frac{\varepsilon_n^2}{2\log_e 2}\}\right| \quad (66)$$

$$\le \left|\{r : \mathrm{TV}\Big(\frac{r}{m_n}, 1-u_1\Big) \le \varepsilon_n\}\right| \quad (67)$$

$$= O(m_n\varepsilon_n) \quad (68)$$

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

$$\Pi_r^{-1} \le \frac{1}{(1-u_1)u_1} \quad (69)$$

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

$$\Pr(A+B=r) = \sum_{i=1}^{r} \Pr(A=r-i)\Pr(B=i) + \Pr(A=r)\Pr(B=0) \quad (70)$$

$$= Q_{1,1}^r Q_{2,2}^{m_n-r} + \sum_{i=1}^{r} \binom{r}{i} Q_{1,1}^{r-i}(1-Q_{1,1})^i \binom{m_n-r}{i} Q_{2,1}^i (1-Q_{2,1})^{m_n-r-i} \quad (71)$$

Again, from Stirling's approximation on the binomial coefficient in (71) and [38, Theorem 11.1.2], we have

$$\Pr(A+B=r) \le Q_{1,1}^r Q_{2,2}^{m_n-r} + \frac{e^2}{2\pi} r^{-1/2}(m_n-r)^{-1/2} U \quad (72)$$

where

$$U = \sum_{i=1}^{r} \Pi_{i/r}^{-1} \Pi_{i/m_n-r}^{-1} 2^{-rD(1-\frac{i}{r}\|Q_{1,1}) - (m_n-r)D(\frac{i}{m_n-r}\|Q_{2,1})} \quad (73)$$

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

$$\Pr(A+B=r) \le Q_{1,1}^r Q_{2,2}^{m_n-r} + \frac{e^2}{2\pi} \frac{{m_n}^{-1}}{\sqrt{(1-u_1-\varepsilon_n)(u_1-\varepsilon_n)}} U \quad (74)$$

and

$$U \le \sum_{i=1}^{r} \Pi_{i/r}^{-1} \Pi_{i/m_n-r}^{-1} 2^{-m_n\left[(1-u_1-\varepsilon_n)D(1-\frac{i}{r}\|Q_{1,1}) + (u_1-\varepsilon_n)D(\frac{i}{m_n-r}\|Q_{2,1})\right]} \quad (75)$$

$$= \sum_{i \notin \mathscr{R}(\varepsilon_n)} \Pi_{i/r}^{-1} \Pi_{i/m_n-r}^{-1} 2^{-m_n\left[(1-u_1-\varepsilon_n)D(1-\frac{i}{r}\|Q_{1,1}) + (u_1-\varepsilon_n)D(\frac{i}{m_n-r}\|Q_{2,1})\right]} + \sum_{i \in \mathscr{R}(\varepsilon_n)} \Pi_{i/r}^{-1} \Pi_{i/m_n-r}^{-1} 2^{-m_n\left[(1-u_1-\varepsilon_n)D(1-\frac{i}{r}\|Q_{1,1}) + (u_1-\varepsilon_n)D(\frac{i}{m_n-r}\|Q_{2,1})\right]} \quad (76)$$

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

$$\mathscr{R}(\varepsilon_n) \triangleq \left\{ i \in [r] : D\Big(1-\frac{i}{r}\Big\|Q_{1,1}\Big), D\Big(\frac{i}{m_n-r}\Big\|Q_{2,1}\Big) \le \frac{\varepsilon_n^2}{2\log_e 2} \right\} \quad (77)$$

Note that similar to T_1, the first summation in (76) 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

$$O(|\mathscr{R}(\varepsilon_n)|) = O(m_n\varepsilon_n) \quad (78)$$

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

$$\Pr(A+B=r) = O(\varepsilon_n) \quad (79)$$

$$T = O(m_n\varepsilon_n^2) = O(V_n^2) \quad (80)$$

$$\mu_n = O(n^2 {m_n}^{-1/2} V_n^2) \quad (81)$$

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

$$\mu_n \le O(W_n^{-1/2} V_n^2) = o(1) \quad (82)$$

Thus $m_n = \omega(n^4)$ is enough to have $\mu_n \to 0$ as $n \to \infty$. □