Electronic Colloquium on Computational Complexity, Report No. 17 (2006)

Improved Lower Bounds for Families of ε-Approximate k-Restricted Min-Wise Independent Permutations

Toshiya Itoh

Global Scientific Information and Computing Center, Tokyo Institute of Technology
2–12–1 O-okayama, Meguro-ku, Tokyo 152-8550, Japan
Tel: +81-3-5734-3651 titoh@dac.gsic.titech.ac.jp

Abstract: A family $\mathcal{F}$ of min-wise independent permutations is known to be a useful tool of indexing replicated documents on the Web. For any integer $n > 0$, let S_n be the family of all permutations on $[1,n] = \{1,2,\ldots,n\}$. For any integer $k \in [1,n]$ and any real $\varepsilon > 0$, we say that a family $\mathcal{F} \subseteq S_n$ of permutations is ε-approximate k-restricted min-wise independent if for any (nonempty) $X \subseteq [1,n]$ such that $\|X\| \leq k$ and any $x \in X$, $|\Pr[\min\{\pi(X)\} = \pi(x)] - 1/\|X\|| \leq \varepsilon/\|X\|$, when π is chosen from $\mathcal{F}$ uniformly at random (where $\|A\|$ is the cardinality of a finite set A). For the size of families $\mathcal{F} \subseteq S_n$ of ε-approximate k-restricted min-wise independent permutations, the following results are known: For any integer $k \in [1,n]$ and any real $\varepsilon > 0$, (constructive upper bound) $\|\mathcal{F}\| = 2^{4k+0(k)} k^{2\log\log(n/\varepsilon)}$; (nonconstructive upper bound) $\|\mathcal{F}\| = O(\frac{k^2}{\varepsilon^2}\log(n/k))$; (lower bound) $\|\mathcal{F}\| = \Omega(k^2(1-\sqrt{8\varepsilon}))$. In this paper, we first derive an upper bound for the Ramsey number of the edge coloring with $m \geq 2$ colors of a complete graph K_ℓ of ℓ vertices, and by the linear algebra method, we then derive a slightly improved lower bound, i.e., we show that for any family $\mathcal{F} \subseteq S_n$ of ε-approximate k-restricted min-wise independent permutations, $\|\mathcal{F}\| = \Omega\left(k\sqrt{\frac{1}{\varepsilon}\log(n/k)}\right)$.

key words: Min-Wise Independence, Positive Definite, Ramsey Number, Linear Algebra Method.

1 Introduction

1.1 Background

The notion of "a family of min-wise independent permutations" was introduced by Broder, et al. [3]. It is a basic tool to estimate *resemblance* between documents [2] and has applications of detecting almost identical documents on the Web [2] and of reducing the amount of randomness used by probabilistic algorithms [9]. Among the several variants of min-wise independence, we focus on the notion of ε-approximate k-restricted min-wise independence.

For any pair of integers $a \leq b$, let $[a,b] = \{a, a+1, \ldots, b\}$, and for any integer $n \geq 1$, we use S_n to denote the family of all permutations on $[1,n]$. For a finite set A, let $\|A\|$ be the *cardinality* of the set A. Informally, we say that a family $\mathcal{F} \subseteq S_n$ of permutations is ε-approximate k-restricted min-wise independent if for any $X \subseteq [1,n]$ such that $\|X\| \leq k$ and any $x \in X$, $\pi(x)$ is the minimum among the images $\pi(X)$ almost equally likely, when π is chosen from $\mathcal{F}$ uniformly at random. More formally,

Definition 1.1 [3]: *For any pair of integers n, k such that $n \geq k \geq 1$ and any real $\varepsilon > 0$, we say that a family $\mathcal{F} \subseteq S_n$ of permutations is ε-*approximate k-restricted min-wise independent* if for any (non-*

ISSN 1433-8092

empty) $X \subseteq [1,n]$ *such that* $\|X\| \leq k$ *and any* $x \in X$*,*

$$\left| \Pr_{\pi \in \mathcal{F}}[\min\{\pi(X)\} = \pi(x)] - \frac{1}{\|X\|} \right| \leq \frac{\varepsilon}{\|X\|}, \tag{1}$$

when π *is chosen from* $\mathcal{F}$ *uniformly at random.*

Let $\mathcal{D}$ be a distribution (not necessarily uniform) on $\mathcal{F}$. We say that a family $\mathcal{F} \subseteq S_n$ of permutations is ε-approximate k-restricted min-wise independent w.r.t. the distribution $\mathcal{D}$ on $\mathcal{F}$ if Equation (1) of Definition 1.1 holds, when π is chosen from $\mathcal{F} \subseteq S_n$ according to the distribution $\mathcal{D}$ on $\mathcal{F}$. For simplicity, we say that a family $\mathcal{F} \subseteq S_n$ of permutations is ε-approximate min-wise independent (resp. k-restricted min-wise independent) if $k = n$ (resp. if $\varepsilon = 0$).

By experiments, a notion of resemblance [2] is known to play an essential role for detection of almost identical documents on the Web. To estimate $r(A,B)$, resemblance between documents A and B, one computes the estimate $\tilde{r}(A,B)$ of the resemblance $r(A,B)$ as follows:

Estimation for the Resemblance $r(A,B)$

(1) Map the document A to the set $D_A \subseteq [1,n]$ by *shingling* [2];

(2) Choose $\pi_1, \pi_2, \ldots, \pi_\ell \in S_n$ independently and uniformly at random;

(3) Define the *sketch* of D_A to be $S_A = (\min\{\pi_1(D_A)\}, \min\{\pi_2(D_A)\}, \ldots, \min\{\pi_\ell(D_A)\})$;

(4) For sketches $S_A = (s_A^1, s_A^2, \ldots, s_A^\ell)$ of A and $S_B = (s_B^1, s_B^2, \ldots, s_B^\ell)$ of B, let

$$\tilde{r}_\ell(A,B) = \frac{\|\{i \in [1,\ell] : s_A^i = s_B^i\}\|}{\ell}.$$

It is easy to see that $\tilde{r}_\ell(A,B)$ converges to $r(A,B)$ quickly when ℓ goes to infinity. Thus $\tilde{r}_\ell(A,B)$ is a good estimation of $r(A,B)$ for finite ℓ and can be used to detect almost identical documents. On the other hand, Broder, et al. [3] showed that any family $\mathcal{F} \subseteq S_n$ of (k-restricted) min-wise independent permutations can be used to compute $r(A,B)$ instead of S_n. Thus in the practical point of view, any family $\mathcal{F} \subseteq S_n$ of ε-approximate k-restricted min-wise independent permutations is an indispensable tool for the detection of almost identical documents.

In this paper, we will investigate the size of families $\mathcal{F} \subseteq S_n$ of ε-approximate k-restricted min-wise independent permutations to precisely capture their inherent nature.

1.2 Known Results

For families $\mathcal{F} \subseteq S_n$ of ε-approximate k-restricted min-wise independent permutations, Broder, et al. [3] showed the following results: (constructive upper bound) $\|\mathcal{F}\| = 2^{4k+0(k)} k^{2\log\log(n/\varepsilon)}$; (nonconstructive upper bound) $\|\mathcal{F}\| = O(\frac{k^2}{\varepsilon^2}\log(n/k))$; (lower bound) $\|\mathcal{F}\| = \Omega(k^2(1-\sqrt{8\varepsilon}))$. It is obvious that there exists a large gap between upper and lower bounds for $\|\mathcal{F}\|$. In particular, the currently known best lower bound for $\|\mathcal{F}\|$ only depends on k and ε but does not depend on n, which seems unnatural. Thus there must exist a tighter lower bound for $\|\mathcal{F}\|$ that depends on n, k, and ε.

The following table summarizes known results on min-wise independent permutations, ε-approximate min-wise independent permutations, k-restricted min-wise independent permutations, etc.

Table 1: Known Results on the Size of Permutation Families

	Upper Bound	Lower Bound
min-wise	uniform: 4^n [3] uniform: $\mathrm{lcm}(n, n-1, \ldots, 1)$ [5] biased: $n2^{n-1}$ [3]	uniform: $\mathrm{lcm}(n, n-1, \ldots, 1)$ [3] biased: $\Omega(\sqrt{n}2^n)$ [3]
ε-approximate min-wise	uniform: $O\left(\frac{n^2}{\varepsilon^2}\right)$ [3] uniform: $n^{O(\log 1/\varepsilon)}$ [4] uniform: $n^{O(\sqrt{\log 1/\varepsilon})}$ [11]	uniform: $n^2(1-\sqrt{8\varepsilon})$ [3] biased: $\max\limits_{r \geq 1} \frac{(n-r)\binom{n}{r}}{1+\varepsilon\binom{n}{r}}$ [3]
k-restricted min-wise	uniform: $(2n)^k \,\mathrm{lcm}(k-2, \ldots, 2)$ [5] uniform: $O(n \lg^2 n)$ $(k=3)$ [12] uniform: $O(n \lg^3 n)$ $(k=4)$ [12] biased: $\sum_{j=1}^{k} j \binom{n}{j}$ [3] biased: $1 + \sum_{j=2}^{k} (j-1)\binom{n}{j}$ [8]	uniform: $\mathrm{lcm}(k, k-1, \ldots, 1)$ [3] uniform: $n-1$ $(k \geq 3)$ [5] biased: $\Omega\left(k2^{k/2} \ln \left(\frac{n}{k}\right)\right)$ [3] biased: $\Omega\left(\binom{n-1}{\lfloor (k-1)/2 \rfloor}\right)$ [10, 6] biased: $1 + \sum_{j=2}^{k} (j-1)\binom{n}{j}$ [8]

1.3 Main Results

In this paper, we will show an improved lower bound for the size of families $\mathcal{F} \subseteq S_n$ of ε-approximate k-restricted min-wise independent permutations. More precisely, we will show (in Theorem 4.1) that for any constant $0 < \varepsilon < 1/5$ and any integer $k \geq 3$, if a family $\mathcal{F} \subseteq S_n$ of permutations is ε-approximate k-restricted min-wise independent for any sufficiently large n, then $\|\mathcal{F}\| = \Omega\left(k\sqrt{\frac{1}{\varepsilon}\log(n/k)}\right)$.

To bound the size of any family $\mathcal{F} \subseteq S_n$ of ε-approximate k-restricted min-wise independent permutations, (i) define $N \times N$ matrices $V_1, V_2, \ldots, V_s$ from the underlying family $\mathcal{F}$ (see Equation (2) in Section 2); (ii) observe that $\|\mathcal{F}\| \geq \mathrm{rank}(V_1) + \mathrm{rank}(V_2) + \cdots + \mathrm{rank}(V_s)$ (see Proposition2.3) by the linear algebra method [7]; (iii) regard V_h as a multi-color edge coloring of a complete graph K_N of N vertices (see Lemma 3.1); (iv) derive lower bounds for $\mathrm{rank}(V_h)$ (see Lemma 3.2).

2 Preliminaries

For any integer $k \geq 3$, let $s = k/3$ and $L = n/s$ (we assume for simplicity that s and L are integers), and partition $[1, n]$ into L disjoint subsets $X_0, X_1, \ldots, X_{L-1}$ of size s, i.e., for each $i \in [0, L-1]$, $X_i = \{si+1, si+2, \ldots, (i+1)s\}$. For any constant $\varepsilon > 0$, let $\mathcal{F} = \{\pi_1, \pi_2, \ldots, \pi_d\} \subseteq S_n$ be a family of ε-approximate k-restricted min-wise independent permutations. Let $N = L-1 = n/s-1$, and for each $h \in [1, s]$, we define an $N \times d$ matrix $U_h = (u_{ij}^h)$ as follows:

$$u_{ij}^h = \begin{cases} \dfrac{1}{\sqrt{d}} & \min\{\pi_j(\{h\} \cup X_0 \cup X_i)\} = \pi_j(h); \\ 0 & \text{otherwise.} \end{cases}$$

For each $h \in [1, s]$, we also define an $N \times N$ matrix $V_h = (v_{ij}^h)$ by the product of U_h and U_h^T, i.e.,

$$V_h = (v_{ij}^h) = U_h U_h^T = \begin{bmatrix} \dfrac{\delta_{11}^h}{2s} & \dfrac{\delta_{12}^h}{3s} & \dfrac{\delta_{13}^h}{3s} & \cdots & \dfrac{\delta_{1N}^h}{3s} \\ \dfrac{\delta_{12}^h}{3s} & \dfrac{\delta_{22}^h}{2s} & \dfrac{\delta_{23}^h}{3s} & \cdots & \dfrac{\delta_{2N}^h}{3s} \\ \dfrac{\delta_{13}^h}{3s} & \dfrac{\delta_{23}^h}{3s} & \dfrac{\delta_{33}^h}{2s} & \cdots & \dfrac{\delta_{3N}^h}{3s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \dfrac{\delta_{1N}^h}{3s} & \dfrac{\delta_{2N}^h}{3s} & \dfrac{\delta_{3N}^h}{3s} & \cdots & \dfrac{\delta_{NN}^h}{2s} \end{bmatrix}. \tag{2}$$

From the assumption that the family $\mathcal{F} \subseteq S_n$ of permutations is ε-approximate k-restricted min-wise independent, it follows that for any $i, j \in [1, N]$, $1 - \varepsilon \leq \delta_{ij}^h \leq 1 + \varepsilon$. So we have the following:

Proposition 2.1: *For the matrix $V_h = (v_{ij}^h)$ given by Equation* (2), *the following holds*: (i) *For any $i \in [1, N]$, $\frac{1-\varepsilon}{2s} \leq v_{ii}^h \leq \frac{1+\varepsilon}{2s}$*; (ii) *For any $i, j \in [1, N]$ such that $i \neq j$, $\frac{1-\varepsilon}{3s} \leq v_{ij}^h \leq \frac{1+\varepsilon}{3s}$* !%

Proposition 2.2: *If a $t \times t$ matrix $A = (a_{ij})$ satisfies that* (C1) *for any $i, j \in [1, t]$ such that $i \neq j$, $a_{ij} = a > 0$ and* (C2) $\min\{a_{11}, a_{22}, \ldots, a_{tt}\} > a$, *then it is nonsingular.*

Proof: Let $\mathbf{1}_t$ be a column vector, all of which entries are 1. Expand the $t \times t$ matrix A as follows:

$$A = \begin{bmatrix} a_{11} & a & a & \cdots & a \\ a & a_{22} & a & \cdots & a \\ a & a & a_{33} & \cdots & a \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a & a & a & \cdots & a_{tt} \end{bmatrix}$$

$$
= \begin{bmatrix} a & a & a & \cdots & a \\ a & a & a & \cdots & a \\ a & a & a & \cdots & a \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a & a & a & \cdots & a \end{bmatrix} + \begin{bmatrix} a_{11}-a & 0 & 0 & \cdots & 0 \\ 0 & a_{22}-a & 0 & \cdots & 0 \\ 0 & 0 & a_{33}-a & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{tt}-a \end{bmatrix}
$$

$$
= a\mathbf{1}_t\mathbf{1}_t^T + \begin{bmatrix} a_{11}-a & 0 & 0 & \cdots & 0 \\ 0 & a_{22}-a & 0 & \cdots & 0 \\ 0 & 0 & a_{33}-a & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{tt}-a \end{bmatrix}. \tag{3}
$$

It is immediate to see that the first term of Equation (3) is positive semidefinite, and from condition (C2), it follows that the second term of Equation (3) is positive definite. So the matrix A is positive definite and thus we have that the matrix A is nonsingular. ■

From the definition of matrices U_h's, it follows that for any $h, g \in [1, s]$ such that $h \neq g$, $U_h U_g^T = 0$. Let U be an $Ns \times d$ matrix, where $U^T = [U_1^T, U_2^T, \ldots, U_s^T]$. Define an $Ns \times Ns$ matrix V by

$$
V = UU^T = \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_s \end{bmatrix} \begin{bmatrix} U_1^T, U_2^T, U_3^T, \ldots, U_s^T \end{bmatrix} = \begin{bmatrix} V_1 & 0 & 0 & \cdots & 0 \\ 0 & V_2 & 0 & \cdots & 0 \\ 0 & 0 & V_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & V_s \end{bmatrix}.
$$

Then it follows that $\mathrm{rank}(V) = \mathrm{rank}(V_1) + \mathrm{rank}(V_2) + \cdots + \mathrm{rank}(V_s)$. Notice that

$$
\mathrm{rank}(V) = \mathrm{rank}(UU^T) \leq \min\{\mathrm{rank}(U), \mathrm{rank}(U^T)\} = \mathrm{rank}(U) \leq \min\{d, Ns\} \leq d.
$$

Thus we have the following proposition that is essential in the subsequent discussions.

Proposition 2.3: $\|\mathcal{F}\| = d \geq \mathrm{rank}(V) = \mathrm{rank}(V_1) + \mathrm{rank}(V_2) + \cdots + \mathrm{rank}(V_s)$.

3 Analysis for rank(V_h)

3.1 Ramsey Number

Let $K_\ell = (V, E)$ be a complete graph of ℓ vertices, and $\mathcal{C}_m = \{c_1, c_2, \ldots, c_m\}$ be a set of m distinct colors. For a complete graph K_ℓ, we use $\chi : E \to \mathcal{C}_m$ to denote an edge coloring of K_ℓ with the set $\mathcal{C}_m$ of m distinct colors. For any integers $t_1, t_2, \ldots, t_m \geq 3$, define $R(t_1, t_2, \ldots, t_m)$ to be a minimum integer ℓ such that for any edge coloring $\chi : E \to \mathcal{C}_m$, there exists a complete subgraph K_{t_i} of K_ℓ, all of which edges are colored by a single color $c_i \in \mathcal{C}_m$. More formally, $R(t_1, t_2, \ldots, t_m)$ is defined to be the minimum integer ℓ that satisfies the following condition: For any edge coloring $\chi : E \to \mathcal{C}_m$ of K_ℓ, there exists an integer $i \in [1,m]$ and a subgraph $K_{t_i} = (V_i, E_i) \subseteq K_\ell$ of t_i vertices such that for each $e \in E_i$, $\chi(e) = c_i$. When $t_1 = t_2 = \cdots = t_m = t$, we simply use $R_m(t)$ to denote $R(t,t,\ldots,t)$.

Notice that $R(t_1, t_2, \ldots, t_m)$ is a generalization of the Ramsey number $R(t_1,t_2)$ of the coloring by two colors [1]. The following lemma is a natural extension of upper bounds for the Ramsey number $R(t_1,t_2)$ of the coloring with two colors [7, the proof of Theorem 27.3].

Lemma 3.1: *For any integer $m \geq 2$ and any integer $t \geq 1$, $R_m(t) \leq m^{mt-(m-1)}$.*

Proof: Let $\ell = m^{mt-(m-1)}$. For a complete graph $K_\ell = (V,E)$, let $V = \{1,2,\ldots,\ell\}$, and fix an edge coloring $\chi : E \to \mathcal{C}_m$ arbitrarily. In the following, we show that there exists an integer $i \in [1,m]$ and a subgraph $K_t = (V_t, E_t) \subseteq K_\ell$ of t vertices such that for each $e \in E_t$, $\chi(e) = c_i$. Let $S_1 = V$ and, for each $j \geq 1$, we construct $S_j \subseteq V$ and $v_j \in S_j$ by iterating the following procedure PARTITION.

Procedure: PARTITION

(1) For the set S_j, choose a vertex $v_j \in S_j$ arbitrarily.

(2) For each $h \in [1,m]$, let $S_j^h = \{v \in S_j - \{v_j\} : \chi((v,v_j)) = c_h\}$. Notice that $S_j^1, S_j^2, \ldots, S_j^m \subseteq S_j - \{v_j\}$ is a partition of $S_j - \{v_j\}$ into m subsets.

(3) Define S_{j+1} to be the largest set among $S_j^1, S_j^2, \ldots, S_j^m$.

For each $j \geq 1$, it is immediate that $\|S_j^1\| + \|S_j^2\| + \cdots + \|S_j^m\| = \|S_j\| - 1$. Thus we have that $\|S_{j+1}\| \geq (\|S_j\|-1)/m$. From the assumption that $S_1 = V$, i.e., $\|S_1\| = \|V\| = m^{mt-(m-1)}$, it follows that a set of vertices $\widetilde{V} = \{v_1, v_2, \ldots, v_{mt-(m-1)}\}$ is chosen when the procedure PARTITION terminates. For each $h \in [1,m]$, let $\widetilde{V}_h = \{v_j \in \widetilde{V} : S_{j+1} = S_j^h\}$, i.e., m subsets $\widetilde{V}_1, \widetilde{V}_2, \ldots, \widetilde{V}_m \subseteq \widetilde{V}$ is a partition of $\widetilde{V}$. From the definition of $\widetilde{V}_h$'s, we have that $mt-(m-1) = \|\widetilde{V}\| = \|\widetilde{V}_1\| + \|\widetilde{V}_2\| + \cdots + \|\widetilde{V}_m\|$. Then there exists an integer $h_* \in [1,m]$ such that $\|\widetilde{V}_{h_*}\| \geq t$. Let $\widetilde{V}_{h_*} = \{v_{h_1}, v_{h_2}, \ldots, v_{h_\tau}\}$, where $\tau \geq t$.

To complete the proof, it suffices to show that for any pair of vertices $v_f, v_g \in \widetilde{V}_{h_*}$, $\chi((v_f,v_g)) = c_{h_*}$. Without loss of generality, assume that $f < g$. From the definition of S_j's, it is obvious that $S_1 \supseteq S_2 \supseteq \cdots \supseteq S_{mt-(m-1)}$, and from the definitions of S_j's, S_j^h's, and $\widetilde{V}_h$'s, it follows that every vertex $v \in S_f^{h_*} = S_{f+1}$ is connected to the vertex v_f with an edge colored by c_{h_*}. Then from the assumption that $f < g$ and the definitions of v_j's, S_j's, and S_j^h's, we immediately have that $v_g \in S_g \subseteq S_{f+1} = S_f^{h_*}$. So it follows that for any pair of vertices $v_f, v_g \in \widetilde{V}_{h_*}$, $\chi((v_f,v_g)) = c_{h_*}$. ■

3.2 Lower Bound for rank(V_h)

To get a stronger lower bound for $\|\mathcal{F}\|$, we need to derive a larger lower bound for rank(V_h). In the subsequent discussions, we show that for each $N \times N$ symmetric matrix V_h in Equation (2), there exists a $t \times t$ submatrix W_h of V_h that satisfies the conditions (C1) and (C2) of Proposition 2.2. This implies have that for each $h \in [1,s]$, rank(V_h) $\geq t$. In fact, we show the following lemma:

Lemma 3.2: *For any constant $0 < \varepsilon < 1/5$ and any integer $k \geq 3$, let $\mathcal{F} \subseteq S_n$ be a family of ε-approximate k-restricted min-wise independent permutations, and for any integer $m \geq 1$, assume that $\|\mathcal{F}\| < \frac{k}{2\varepsilon}m$. Then for each $h \in [1,s]$, $\mathrm{rank}(V_h) = N$ if $m = 1$; $\mathrm{rank}(V_h) \geq \lfloor \frac{\log(3n/k)}{m \log m} \rfloor$ if $m \geq 2$.*

3.2.1 Intuition Behind the Proof of Lemma 3.2

To see the intuition behind the proof of Lemma 3.2, let us consider the following simple case: Assume that for any integer $m \geq 1$, $\|\mathcal{F}\| < \frac{k}{2\varepsilon}m$. This implies that every offdiagonal entry of the $N \times N$ symmetric matrix V_h given by Equation (2) can take at most m values. We regard these m values as the edge coloring of a complete graph $K_N = (V, E)$ of N vertices with m colors. So from Lemma 3.1, we have that if $N \geq R_m(t)$, then there exists a subgraph $K_t = (V_t, E_t)$ of t vertices, all of which edges are colored with a single color, which guarantees that there exists a $t \times t$ submatrix W_h of V_h satisfying the condition (C1) of Proposition 2.2. From the assumption that $0 < \varepsilon < 1/5$, we also have that the submatrix W_h of V_h satisfies the condition (C2) of Proposition 2.2. Then from Proposition 2.2, it follows that the $t \times t$ submatrix W_h of V_h is nonsingular and thus $\mathrm{rank}(V_h) \geq \mathrm{rank}(W_h) = t$.

3.2.2 Proof of Lemma 3.2

For any $n \geq 3$ and any $k \in [3, n]$, recall that $s = k/3$ and $N = n/s - 1$. Then for each $h \in [1, s]$ and any pair of $i, j \in [1, N]$ such that $i \neq j$, we define a subfamily $\mathcal{G}_{ij}^h \subseteq \mathcal{F}$ of permutations by

$$\mathcal{G}_{ij}^h = \{\pi \in \mathcal{F} : \min\{\pi(\{h\} \cup X_0 \cup X_i \cup X_j)\} = \pi(h)\},$$

and let $G_{ij}^h = \|\mathcal{G}_{ij}^h\|$. Since the family $\mathcal{F} = \{\pi_1, \pi_2, \ldots, \pi_d\} \subseteq S_n$ of permutations is ε-approximate k-restricted min-wise independent, we have that $\frac{1-\varepsilon}{3s} \leq G_{ij}^h / \|\mathcal{F}\| \leq \frac{1+\varepsilon}{3s}$. Then it follows that

$$\frac{1-\varepsilon}{k}\|\mathcal{F}\| \leq G_{ij}^h \leq \frac{1+\varepsilon}{k}\|\mathcal{F}\|. \tag{4}$$

Since G_{ij}^h is an positive integer, it is obvious that for any integer $m \geq 1$, if $\frac{2\varepsilon}{k}\|\mathcal{F}\| < m$, then G_{ij}^h can take at most m possible integers. This implies that for each $h \in [1, s]$, every offdiagonal entry of the matrix V_h given by Equation (2) is restricted to m possible values. Since the matrix V_h is symmetric, we regard V_h as the adjacency matrix of a complete graph K_N and regard these m values of the offdiagonal entries of V_h as edge coloring of K_N with the set $\mathcal{C}_m = \{c_1, c_2, \ldots, c_m\}$ of m colors. Let us consider the following cases: (Case 1) $m = 1$; (Case 2) $m \geq 2$.

(Case 1) Since $m = 1$, we have that for each $h \in [1, s]$, every offdiagonal entry of the $N \times N$ symmetric matrix $V_h = (v_{ij}^h)$ given by Equation (2) is restricted to a single value v. Then the matrix V_h satisfies the condition (C1) of Proposition 2.2. From Proposition 2.1, it is immediate to see that for each $h \in [1, s]$, $\min\{v_{11}^h, v_{22}^h, \ldots, v_{NN}^h\} \geq \frac{1-\varepsilon}{2s}$ and $\frac{1-\varepsilon}{3s} \leq v \leq \frac{1+\varepsilon}{3s}$. Thus from the assumption that $0 < \varepsilon < 1/5$, it follows that $\min\{v_{11}^h, v_{22}^h, \ldots, v_{NN}^h\} > v > 0$, which implies that the matrix V_h satisfies the condition (C2) of Proposition 2.2. So for each $h \in [1, s]$, V_h is nonsingular, i.e., $\mathrm{rank}(V_h) = N$.

(Case 2) Since $m \geq 2$, we have that for each $h \in [1, s]$, every offdiagonal entry of the $N \times N$ symmetric matrix $V_h = (v_{ij}^h)$ given by Equation (2) is restricted to $m \geq 2$ values. We regard V_h as the adjacency matrix of a complete graph K_N of N vertices and also regard these m values of the offdiagonal entries of V_h as edge coloring of K_N with the set $\mathcal{C}_m = \{c_1, c_2, \ldots, c_m\}$ of $m \geq 2$ colors. It follows from Lemma 3.1 that for any coloring $\chi : E \to \mathcal{C}_m$ of a complete graph $K_\ell = (V, E)$ of ℓ vertices such that $\|V\| = \ell \geq m^{mt-(m-1)}$, there exists an integer $i \in [1, m]$ and a subgraph $K_t = (V_t, E_t) \subseteq K_\ell$ of t vertices such that for each $e \in E_t$, $\chi(e) = c_i$. So for any integer $m \geq 2$, if $N = \frac{3n}{k} - 1 \geq m^{mt-(m-1)}$, then for each $h \in [1, s]$, there exists a $t \times t$ submatrix $W_h = (w_{ij}^h)$ of V_h that satisfies the condition (C1) of Proposition 2.2. In a way similar to (Case 1), we can show that the submatrix W_h satisfies the condition (C2) of Proposition 2.2. Then for each $h \in [1, s]$, W_h is nonsingular, i.e., $\mathrm{rank}(W_h) \geq t$. To guarantee that $\frac{3n}{k} - 1 \geq m^{mt-(m-1)}$, we can take any integer $t \geq 1$ such that $\frac{3n}{k} \geq m^{mt}$, and let $t_* = \lfloor \frac{\log(3n/k)}{m \log m} \rfloor$ be the maximum among those values of $t \geq 1$. Thus we have that for each $h \in [1, s]$,

$$\mathrm{rank}(V_h) \geq \mathrm{rank}(W_h) \geq t_* = \left\lfloor \frac{\log(3n/k)}{m \log m} \right\rfloor .$$

4 Main Result

In this section, we derive a main result of the paper, i.e., a lower bound for the size of families $\mathcal{F} \subseteq S_n$ of ε-approximate k-restricted min-wise independent permutations. In fact, we show the following:

Theorem 4.1: *For any constant $0 < \varepsilon < 1/5$ and any integer $k \geq 3$, if a family $\mathcal{F} \subseteq S_n$ of permutations is ε-approximate k-restricted min-wise independent for any sufficiently large n, then*

$$\|\mathcal{F}\| = \Omega\left(k \cdot \frac{\log^{1/2-o(1)}(n/k)}{\varepsilon^{1/2+o(1)}}\right).$$

4.1 Intuition Behind the Proof of Theorem 4.1

The proof of Theorem 4.1 is based on the following observation: For any integer $m \geq 1$, let us consider the case that $\|\mathcal{F}\| < \frac{k}{2\varepsilon}m$, which implies that the underlying family $\mathcal{F} \subseteq S_n$ of permutations is small. Assume that $m = 1$. So from Lemma 3.2, we have that for each $h \in [1, s]$, $\mathrm{rank}(V_h) = N$, and from Proposition 2.3, it follows that $\|\mathcal{F}\| \geq \mathrm{rank}(V) = sN = \frac{k}{3}(\frac{n}{k} - 1) = \frac{n-k}{3}$, which implies that the underlying family $\mathcal{F} \subseteq S_n$ of permutations is large. For sufficiently large n's, this contradicts the assumption that $\|\mathcal{F}\| < \frac{k}{2\varepsilon}$, and thus we have that $\|\mathcal{F}\| \geq \frac{k}{2\varepsilon}$. In a way similar to the observation above, we can use Lemma 3.2 and Proposition 2.3 to show that if n is sufficiently large, then for any integer $m \in [1, m_*]$, $\|\mathcal{F}\| \geq \frac{k}{2\varepsilon}m$. Thus for any sufficiently large n, we will determine the value of $m_* \geq 1$ as large as possible to derive a larger lower bound for $\|\mathcal{F}\|$.

4.2 Proof of Theorem 4.1

We show the theorem more formally. For any integer $m \geq 1$, assume that $\|\mathcal{F}\| < \frac{k}{2\varepsilon}m$. Then for any constant $0 < \varepsilon < 1/5$ and any integer $k \geq 3$, it immediately follows from Lemma 3.2 that for each $h \in [1, s]$, $\mathrm{rank}(V_h) \geq \lfloor \frac{\log(3n/k)}{m \log m} \rfloor$. So from Proposition 2.3, we have that

$$\begin{aligned} \|\mathcal{F}\| &\geq \mathrm{rank}(V) = \mathrm{rank}(V_1) + \mathrm{rank}(V_2) + \cdots + \mathrm{rank}(V_s) \\ &\geq s \left\lfloor \frac{\log(3n/k)}{m \log m} \right\rfloor = \frac{k}{3} \left\lfloor \frac{\log(3n/k)}{m \log m} \right\rfloor . \end{aligned}$$

If there exists an integer $m \geq 1$ such that $\frac{k}{2\varepsilon}m \leq \frac{k}{3}\lfloor\frac{\log(3n/k)}{m\log m}\rfloor$, then $\frac{k}{2\varepsilon}m \leq \|\mathcal{F}\|$, which contradicts the assumption that $\|\mathcal{F}\| < \frac{k}{2\varepsilon}m$. It is easy to see that for any constant $0 < \varepsilon < 1/5$ and any integer $k \geq 3$, if n is sufficiently large, then there always exist integers m's such that $\frac{k}{2\varepsilon}m \leq \frac{k}{3}\lfloor\frac{\log(3n/k)}{m\log m}\rfloor$. Thus for any sufficiently large n, we have that $\|\mathcal{F}\| \geq \frac{k}{2\varepsilon}m$ for any integer $m \geq 1$ such that $\frac{k}{2\varepsilon}m \leq \frac{k}{3}\lfloor\frac{\log(3n/k)}{m\log m}\rfloor$.

To achieve a larger lower bound for $\|\mathcal{F}\|$, it suffices to take the maximum m_* among those integers m's such that $\frac{k}{2\varepsilon}m \leq \frac{k}{3}\lfloor\frac{\log(3n/k)}{m\log m}\rfloor$. Since the maximum m_* satisfies that

$$\frac{2\varepsilon}{3}\log(3n/k) \geq m_*^2\log m_* = m_*^{2+o(1)},$$

we have that for any constant $0 < \varepsilon < 1/5$ and any integer $k \geq 3$, $m_* = \{\frac{2\varepsilon}{3}\log(3n/k)\}^{1/2-o(1)}$ (if n is sufficiently large). Thus for any constant $0 < \varepsilon < 1/5$ and any integer $k \geq 3$, if a family $\mathcal{F} \subseteq S_n$ of permutations is ε-approximate k-restricted min-wise independent for any sufficiently large n, then

$$\|\mathcal{F}\| \geq \frac{k}{2\varepsilon}m_* = \Omega\left(k \cdot \frac{\log^{1/2-o(1)}(n/k)}{\varepsilon^{1/2+o(1)}}\right).$$

5 Concluding Remarks

In this paper, we have derived an improved lower bound for the size of families $\mathcal{F} \subseteq S_n$ of ε-approximate k-restricted min-wise independent permutations, i.e., we have shown (in Theorem 4.1) that for any constant $0 < \varepsilon < 1/5$ and any integer $k \geq 3$, if a family $\mathcal{F} \subseteq S_n$ of permutations is ε-approximate k-restricted min-wise independent for any sufficiently large n, then $\|\mathcal{F}\| = \Omega\left(k\sqrt{\frac{1}{\varepsilon}\log(n/k)}\right)$.

The result of Theorem 4.1 is based on the matrix formulations of the underlying family $\mathcal{F} \subseteq S_n$ of ε-approximate k-restricted min-wise independent permutations and the multi-color edge coloring of a complete graph K_ℓ of ℓ vertices. For the multi-color edge coloring of K_ℓ, we have regarded it as the Ramsey number [7] and have derived its upper bound for the size of K_ℓ (see Lemma 3.1). This is the main observation to derive Theorem 4.1 and would be of independent interest.

As for the size of families $\mathcal{F} \subseteq S_n$ of ε-approximate k-restricted min-wise independent permutations, the currently known best upper bound is $\|\mathcal{F}\| = O(\frac{k^2}{\varepsilon^2}\log(n/k))$ due to Broder, et al. [3], and our lower bound given by Theorem 4.1 still has a gap to the best upper bound. Then for any family $\mathcal{F} \subseteq S_n$ of ε-approximate k-restricted min-wise independent permutations,

(1) derive tight upper and lower bounds for $\|\mathcal{F}\|$.

For any family $\mathcal{F} \subseteq S_n$ of k-restricted min-wise independent permutations, we have already known that $\|\mathcal{F}\| = \Omega\left(n^{\lfloor(k-1)/2\rfloor}\right)$ for any distribution $\mathcal{D}$ on $\mathcal{F}$ [10, 6]. On the other hand, for any family $\mathcal{F} \subseteq S_n$ of ε-approximate k-restricted min-wise independent permutations, Theorem 4.1 holds only for the uniform distribution $\mathcal{U}$ on $\mathcal{F}$ (and it would be hard to extend to the case of *biased* distribution $\mathcal{D}$ on $\mathcal{F}$). Thus for any family $\mathcal{F} \subseteq S_n$ of ε-approximate k-restricted min-wise independent permutations,

(2) derive a lower bound for $\|\mathcal{F}\|$ w.r.t. any distribution $\mathcal{D}$ on $\mathcal{F}$.

References

[1] Alon, N. and Spencer, J., The Probabilistic Method, John Wiley & Sons, 1992.

[2] Broder, A., On the Resemblance and Containment of Documents, in *Proc. of Compression and Complexity of Sequences*, 21–29, 1998.

[3] Broder, A., Charikar, M., Frieze, A., and Mitzenmacher, M., Min-Wise Independent Permutations, *J. Comput. Sys. Sci.*, 60(3):630–659, 2000. A preliminary version was appeared in *Proc. of the 30th Annual ACM Symposium on Theory of Computing*, 327–336, 1998.

[4] Indyk, P., A Small Approximately Min-Wise Independent Family of Hash Functions, *J. of Algorithms*, 38:84–90, 2001. A preliminary version was appeared in *Proc. of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms*, 454–456, 1999.

[5] Itoh, T., Takei, Y., and Tarui, J., On Permutations with Limited Independence, in *Proc. of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms*, 137–146, 2000.

[6] Itoh, T., Takei, Y., and Tarui, J., On the Sample Size k-Restricted Min-Wise Independent Permutations and Other k-Wise Distributions, in *Proc. of the 35th Annual ACM Symposium on Theory of Computing*, 710–719, 2003.

[7] Jukna, S., Extremal Combinatorics, Springer, 2001.

[8] Matoušek, J. and Stojaković, M., On Restricted Min-Wise Independence of Permutations, *Preprint*, 2002. Available at http://kam.mff.cuni.cz/~matousek/preprints.html/

[9] Mulmuley, K., Randomized Geometric Algorithms and Pseudorandom Generators, *Algorithmica*, 16:450–463, 1996.

[10] Norin, S., A Polynomial Lower Bound for the Size of any k-Min-Wise Independent Set of Permutation, *Zapiski Nauchnyh Seminarov (POMI)*, 277:104–116, 2001 (in Russian). Available at http://www.pdmi.ras.ru/znsl/

[11] Saks, M., Srinivasan, A., Zhou, S., and Zuckerman, D., Low Discrepancy Sets Yield Approximate Min-Wise Independent Permutation Families, *Inform. Process. Lett.*, 73:29–32, 2000. A preliminary version was appeared in *Proc. of RANDOM–APPROX'99*, Lecture Notes in Computer Science 1671, Springer, 11–15, 1999.

[12] Tarui, J., Itoh, T., and Takei, Y., A Nearly Linear Size 4-Min-Wise Independent Permutation Family by Finite Geometries, in *Proc. of RANDOM–APPROX'03*, Lecture Notes in Computer Science 2764, Springer, 396–408, 2003.