*Completeness and Robustness Properties of Min-Wise Independent Permutations**

Andrei Z. Broder[1], Michael Mitzenmacher[2]
[1] *AltaVista Company, 1825 S. Grant Street, Suite 410, San Mateo, California 94402; e-mail: andrei.broder@altavista.com*
[2,†] *Harvard University, 33 Oxford St., Cambridge, Massachusetts 02138; e-mail: michaelm@eecs.harvard.edu*

Received 2 December 1999; accepted 1 June 2000

ABSTRACT: We provide several new results related to the concept of min-wise independence. Our main result is that any randomized sampling scheme for the relative intersection of sets based on testing equality of samples yields an equivalent min-wise independent family. Thus, in a certain sense, min-wise independent families are "complete" for this type of estimation. We also discuss the notion of robustness, a concept extending min-wise independence to allow more efficient use of it in practice. A surprising result arising from our consideration of robustness is that under a random permutation from a min-wise independent family, any element of a fixed set has an equal chance to get any rank in the image of the set, not only the minimum as required by definition. © 2001 John Wiley & Sons, Inc. Random Struct. Alg., 18, 18–30, 2001

Correspondence to: M. Mitzenmacher.
A preliminary version of this article appeared in the Proceedings of Random-Approx '99.
†Supported in part by the Alfred P. Sloan Foundation Research Fellowship and equipment from the Compaq Computer Corporation. Parts of this work were done while both authors were at Compaq Systems Research Center.

1. INTRODUCTION

A family of permutations $\mathcal{P} \subseteq S_n$ is called *min-wise independent* (abbreviated MWI) if for any set $X \subseteq [n] = \{1, \ldots, n\}$ and any $x \in X$, when π is chosen at random in $\mathcal{P}$ according to some specified probability distribution, we have

$$\mathbf{Pr}\big(\min\{\pi(X)\} = \pi(x)\big) = \frac{1}{|X|} \ . \tag{1}$$

In other words we require that all the elements of any fixed set X have an equal chance to become the minimum element of the image of X under π.

When the distribution on $\mathcal{P}$ is nonuniform, the family is called *biased*, and it is called *unbiased* otherwise. In general in this paper we will not specify the probability distribution on $\mathcal{P}$ unless relevant, and from now on when we say "π chosen at random in (the min-wise independent family) $\mathcal{P}$" we mean "π chosen in $\mathcal{P}$ according to the probability distribution associated to $\mathcal{P}$ such that (1) holds."

Together with Moses Charikar and Alan Frieze, we introduced this notion in [5], motivated by the fact that such a family (under some relaxations) is essential to the algorithm used in practice by the AltaVista web index software to detect and filter near-duplicate documents. The crucial property that enables this application is the following: let X be a subset of $[n]$. Pick a "sample" $s(X) \in X$ by choosing at random a permutation π from a family of permutations $\mathcal{P}$ and letting

$$s(X) = \pi^{-1}(\min\{\pi(X)\}) \ . \tag{2}$$

Then, if $\mathcal{P}$ is a MWI-family, for any two nonempty subsets A and B, we have

$$\mathbf{Pr}\big(s(A) = s(B)\big) = \frac{|A \cap B|}{|A \cup B|} \ . \tag{3}$$

Hence, such samples can be used to estimate the relative size of the intersection of sets, a quantity that we call the *resemblance* of A and B, defined as

$$R(A, B) = \frac{|A \cap B|}{|A \cup B|} \ . \tag{4}$$

We estimate resemblance by first picking, say, 100 permutations from a MWI-family, and then computing samples for each set of interest. Then the resemblance of any two sets can be estimated simply by determining the fraction of samples that coincide.

In practice we can allow small relative errors. We say that $\mathcal{P} \subseteq S_n$ is *approximately min-wise independent with relative error* ϵ (or just approximately min-wise independent, where the meaning is clear) if for any set $X \subseteq [n]$ and any $x \in X$, when π is chosen at random in $\mathcal{P}$ we have

$$\left|\mathbf{Pr}\big(\min\{\pi(X)\} = \pi(x)\big) - \frac{1}{|X|}\right| \le \frac{\epsilon}{|X|} \ . \tag{5}$$

For further details about the use of these ideas to estimate document similarity see [7, 2, 3]. Takei, Itoh, and Shinozaki [14] presented an optimal (size-wise) construction for a MWI-family under the uniform distribution. Their family has size $\mathrm{lcm}(1, \ldots, n)$, matching the lower bound of [5]. Explicit constructions of approx-

imately MWI-families were obtained by Indyk [9] and by Saks et al. [13]. For an application of these families to derandomization see [6].

We also note that concepts similar to min-wise independence have appeared prior to our work [5] as well. For example, the monotone ranged hash functions described in [10] have the min-wise independence property; Cohen [8] uses the property that the minimum element of a random permutation is uniform to estimate the size of the transitive closure, as well as to solve similar related problems; and Mulmuley [12] uses what we call approximate min-wise independence to use fewer random bits for several randomized geometric algorithms.

The main result of this paper, presented in Section 2, is that, rather surprisingly, *any* sampling scheme that has property (3) is equivalent to a scheme derived via Eq. (2) from a min-wise independent family of permutations. More precisely we have the following theorem:

Theorem 1. *Let $\mathcal{F}$ be a family of functions from nonempty subsets of $[n]$ to some arbitrary set Ω. Assume there exists a probability distribution on $\mathcal{F}$ such that when f is chosen from $\mathcal{F}$ according to this distribution, for any two nonempty subsets A and B,*

$$\mathbf{Pr}\big(f(A) = f(B)\big) = \frac{|A \cap B|}{|A \cup B|} .$$

Then there exists a min-wise independent family of permutations $\mathcal{P}$ and a bijection from $\mathcal{F}$ to $\mathcal{P}$ such that every $f \in \mathcal{F}$ is defined by

$$f(X) = f\left(\left\{\pi_f^{-1}(\min\{\pi_f(X)\})\right\}\right)$$

for some $\pi_f \in \mathcal{P}$.

We note here some immediate consequences of the theorem:

- **(a)** The induced family of permutations has the same size as the initial family of functions, that is $|\mathcal{P}| = |\mathcal{F}|$.
- **(b)** Each $f \in \mathcal{F}$ takes exactly n distinct values $f(\{x_1\}), \ldots, f(\{x_n\})$. (*A priori* each f can take $2^n - 1$ values.)
- **(c)** Assume that we add the condition that for every $X \subseteq [n]$, each $f \in \mathcal{F}$ satisfies $f(X) \in X$; in other words, the "sample" must belong to the set being sampled. Then for every $x \in [n]$ each f satisfies $f(\{x\}) = x$, and hence each f has the form

  $$f(X) = \pi_f^{-1}(\min\{\pi_f(X)\}) .$$

  (The converse of the assumption is also true: if for every $x \in [n]$ we have $f(\{x\}) = x$ then $f(X) \in X$ follows. See Lemma 3 below.)
- **(d)** Thus every estimation scheme that has property (3) is equivalent under renaming to a sampling scheme derived via Eq. (2) from a min-wise independent family of permutations. (For each f, $f(\{x_1\})$ is the "name" of x_1, $f(\{x_2\})$ is the "name" of x_2, etc.)

Of course in practice it might be more convenient to represent $\mathcal{F}$ directly rather than via $\mathcal{P}$. (See [4] for an example.) But the fact remains that any method of

sampling to estimate resemblance via Eq. (3) is equivalent to sampling with min-wise independent permutations.

To develop some intuition, before plunging into the proof, we start by observing that the choice of "min" in the definition (1) is somewhat arbitrary. Clearly if we replace "min" with "max" both in (1) and in (2), property (3) holds. More generally, we can fix a permutation $\sigma \in S_n$ (think of it as a total order on $[n]$), and require $\mathcal{P}$ to satisfy the property

$$\mathbf{Pr}\big(\min\{\sigma(\pi(X))\} = \sigma(\pi(x))\big) = \frac{1}{|X|} \ . \tag{6}$$

Then we can choose samples according to the rule

$$s(X) = \pi^{-1}\left(\sigma^{-1}\left(\min\{\sigma(\pi(X))\}\right)\right) \ .$$

(We obtain "max" by taking $\sigma(i) = n + 1 - i$.)

Is there any advantage to choosing a particular σ? A moment of reflection indicates that there is nothing to be gained since we can simply replace the family $\mathcal{P}$ by changing every $\pi(\cdot) \in \mathcal{P}$ to $\sigma(\pi(\cdot))$. This is, in fact, a very simple instance of Theorem 1. However, it could be of interest if a family $\mathcal{P}$ satisfies condition (6) with respect to more than one order σ. One reason is that, in practice, computing $\pi(X)$ is expensive (see [4] for details). If a family has the min-wise independence property with respect to several orders, then we can extract a sample for each order. Obviously these samples are correlated, but if the correlation can be bounded, these samples are still usable.

Takei, Itoh, and Shinozaki [14] observed that their construction for a MWI-family of size $\mathrm{lcm}(1, \ldots, n)$ under the uniform distribution yields a family that is simultaneously min-wise independent and max-wise independent. In Section 3 we show that this is not a fluke; in fact, any min-wise independent family is also max-wise independent. Moreover, if $\mathcal{P} \subseteq S_n$ is min-wise independent, then for any set $X \subseteq [n]$, any $x \in X$, and any fixed $r \in \{1, \ldots, |X|\}$, when π is chosen at random in $\mathcal{P}$ we have

$$\mathbf{Pr}\big(\mathrm{rank}(\pi(x), \pi(X)) = r\big) = \frac{1}{|X|} \ , \tag{7}$$

where $\mathrm{rank}(x, X)$ for $x \in X$ is the number of elements in X not greater than x. Hence the max-wise independence property follows by taking $r = |X|$.

In Section 4 we discuss families that have the min-wise independence property with respect to *all* possible orders σ. We call such families *robust*. We show that although not every min-wise independent family is robust, there are nontrivial robust families. On the other hand, robust families under the uniform distribution of size $\mathrm{lcm}(1, \ldots, n)$ do not necessarily exist for every n.

2. ANY SAMPLING SCHEME IS A MWI-FAMILY

In this section we prove the following:

Theorem 1. *Let $\mathcal{F}$ be a family of functions from nonempty subsets of $[n]$ to some arbitrary set Ω. Assume there exists a probability distribution on $\mathcal{F}$ such that when f is*

chosen from $\mathcal{F}$ according to this distribution, for any two nonempty subsets A and B,

$$\mathbf{Pr}\big(f(A) = f(B)\big) = \frac{|A \cap B|}{|A \cup B|} .$$

Then there exists a min-wise independent family of permutations $\mathcal{P}$ and a bijection from $\mathcal{F}$ to $\mathcal{P}$ such that every $f \in \mathcal{F}$ is defined by

$$f(X) = f\left(\left\{\pi_f^{-1}(\min\{\pi_f(X)\})\right\}\right)$$

for some $\pi_f \in \mathcal{P}$.

Proof. Assume the premises of the theorem. We start with four simple lemmas.

Lemma 1. *Let X be a nonempty subset of $[n]$. Then for any $x \in X$*

$$\mathbf{Pr}(f(X) = f(\{x\})) = \frac{\big|X \cap \{x\}\big|}{\big|X \cup \{x\}\big|} = \frac{1}{|X|} .$$

Lemma 2. *For any two distinct elements $x_1, x_2 \in [n]$ and each $f \in \mathcal{F}$.*

$$f(\{x_1\}) \neq f(\{x_2\}) .$$

Proof. By hypothesis

$$\mathbf{Pr}(f(\{x_1\}) = f(\{x_2\})) = \frac{\big|\{x_1\} \cap \{x_2\}\big|}{\big|\{x_1\} \cup \{x_2\}\big|} = 0.$$

■

Lemma 3. *Let $X = \{x_1, x_2, \ldots, x_k\}$ be a nonempty subset of $[n]$. Then for each $f \in \mathcal{F}$*

$$f(X) \in \{f(\{x_1\}), f(\{x_2\}), \ldots, f(\{x_k\})\} .$$

Proof.

$$\mathbf{Pr}\left(f(X) \in \{f(\{x_1\}), f(\{x_2\}), \ldots, f(\{x_k\})\}\right)$$

$$= \sum_{i=1}^{k} \mathbf{Pr}\left(f(X) = f(\{x_i\})\right) = 1.$$

■

Lemma 4. *Let $X = \{x_1, x_2, \ldots, x_k\}$ and Y be nonempty subsets of $[n]$. If $X \subseteq Y$, then for every $f \in \mathcal{F}$, if $f(Y) \in \{f(\{x_1\}), f(\{x_2\}), \ldots, f(\{x_k\})\}$, then $f(Y) = f(X)$.*

Proof. By hypothesis

$$\mathbf{Pr}(f(X) = f(Y)) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{k}{|Y|} .$$

On the other hand,

$$\begin{aligned}
&\mathbf{Pr}(f(X) = f(Y)) \\
&\quad = \mathbf{Pr}(f(X) = f(\{x_1\}) \wedge f(Y) = f(\{x_1\})) + \cdots \\
&\qquad + \mathbf{Pr}(f(X) = f(\{x_k\}) \wedge f(Y) = f(\{x_k\})) \\
&\quad = \mathbf{Pr}(f(X) = f(\{x_1\}) \mid f(Y) = f(\{x_1\}))\,\mathbf{Pr}(f(Y) = f(\{x_1\})) + \cdots \\
&\qquad + \mathbf{Pr}(f(X) = f(\{x_k\}) \mid f(Y) = f(\{x_k\}))\,\mathbf{Pr}(f(Y) = f(\{x_k\})) \\
&\quad = \mathbf{Pr}(f(X) = f(\{x_1\}) \mid f(Y) = f(\{x_1\}))(1/|Y|) + \cdots \\
&\qquad + \mathbf{Pr}(f(X) = f(\{x_k\}) \mid f(Y) = f(\{x_k\}))(1/|Y|).
\end{aligned}$$

(The last equality follows from Lemma 1.) Hence for every $x_i \in X$

$$\mathbf{Pr}(f(X) = f(\{x_i\}) | f(Y) = f(\{x_i\})) = 1 ,$$

and therefore for every $f \in F$, if $f(Y) = f(\{x_i\})$ then $f(X) = f(\{x_i\})$ as well. ■

Returning to the proof of the theorem, we show now how to construct for each $f \in \mathcal{F}$ a permutation π_f such that for every nonempty set X

$$f(X) = f\left(\left\{\pi_f^{-1}(\min\{\pi_f(X)\})\right\}\right) . \tag{8}$$

Note that the family $\mathcal{P}$ given by the π_f above is clearly min-wise independent by Lemma 1, as for any $x \in X$,

$$\mathbf{Pr}\left(\min\{\pi_f(X)\} = \pi_f(x)\right) = \mathbf{Pr}(f(X) = f(\{x\})) = \frac{1}{|X|} .$$

Fix f and let $g\colon \{f(\{x_1\}), \ldots, f(\{x_n\})\} \to [n]$ be the function defined by $g(f(\{x_i\})) = x_i$. In view of Lemma 2 g is well-defined. Now define a sequence $y_1, y_2, \ldots, y_n$ as follows:

$$\begin{aligned}
y_1 &= g(f([n])) \\
y_2 &= g(f([n] \setminus \{y_1\})) \\
y_3 &= g(f([n] \setminus \{y_1, y_2\})) \\
&\vdots
\end{aligned}$$

In view of Lemma 3 g is correctly used and we have

$$\begin{aligned} f([n]) &= f(\{y_1\}) \\ f([n] \setminus \{y_1\}) &= f(\{y_2\}) \\ f([n] \setminus \{y_1, y_2\}) &= f(\{y_3\}) \\ &\vdots \end{aligned}$$

Furthermore $y_1, y_2, \ldots, y_n$ is a permutation of $[n]$. Finally we take π_f to be the inverse of the permutation determined by the y_i; that is, π_f maps y_1 to 1, y_2 to 2, etc. We need to show that f satisfies Eq. (8) for every nonempty set X.

Fix X and consider the sets $Y_1 = [n]$, $Y_2 = [n] \setminus \{y_1\}$, $Y_3 = [n] \setminus \{y_1, y_2\}$, $\ldots$, $Y_n = \{y_n\}$. Let k be the largest index such that Y_k still includes X. This implies that

(a) $y_k \in X$ since otherwise we could have taken Y_{k+1}.
(b) $\{y_1, y_2, \ldots, y_{k-1}\} \cap X = \varnothing$ since none of these elements belong to Y_k.

By definition $f(Y_k) = f(\{y_k\})$. But $y_k \in X \subseteq Y_k$ and therefore Lemma 4 implies that $f(X) = f(\{y_k\})$ as well. On the other hand property (a) above implies that $\min\{\pi_f(X)\} \leq k$ and property (b) implies that $\min\{\pi_f(X)\} > k - 1$. Hence $\min\{\pi_f(X)\} = k$ and $\pi_f^{-1}(\min\{\pi_f(X)\}) = y_k$ as required. ■

3. RANK UNIFORMITY FOR MWI-FAMILIES

In this section, we show that any min-wise independent family actually has the property that every item in any fixed set is equally likely to have any rank in the image of the set—not just the minimum rank as required by definition. Our analysis is based on the following lemma, which follows from Theorem 7 of [5].

Lemma 5. *A family of permutations $\mathcal{P}$ is min-wise independent if and only if for any set $X \subset [n]$ of size k and any element $x \in [n] \setminus X$,*

$$\mathbf{Pr}\,(\pi(X) = [k] \wedge \pi(x) = k + 1) = \frac{1}{\binom{n}{k}(n-k)},$$

when π is chosen at random in $\mathcal{P}$.

In other words, if we fix a set X of size k and an extra element x, the probability that x maps to $k+1$ and X maps to $\{1, \ldots, k\}$ in some arbitrary order is exactly what "it should be" if we were sampling uniformly from the entire set of permutations S_n.

Theorem 2. *If $\mathcal{P}$ is min-wise independent, and π is chosen at random from $\mathcal{P}$, then for any set $X \subset [n]$ and any element $x \in X$,*

$$\mathbf{Pr}\big(\text{rank}(\pi(x), \pi(X)) = r\big) = \frac{1}{|X|} . \tag{9}$$

Proof. We sum over all the possible ways such that $\text{rank}(\pi(x), \pi(X)) = r$ and $\pi(x) = s$ and consider which elements map to $[s-1]$. Note that we must have $r \le s \le n - (|X| - r)$. There must be $r-1$ other elements of X, call them $\{x_1, x_2, \ldots, x_{r-1}\}$, such that $\pi(x_i) \in [s-1]$, and there are $\binom{|X|-1}{r-1}$ ways to choose them. Similarly, there must be $s-r$ elements of $[n] \setminus X$, call them $\{y_1, y_2, \ldots, y_{s-r}\}$, such that $\pi(y_i) \in [s-1]$ and there are $\binom{n-|X|}{s-r}$ ways to choose these elements. For each possible combination of choices, we have from Lemma 5 that the probability that these elements are mapped to $[s-1]$ and x is mapped to s is

$$\frac{1}{\binom{n}{s-1}(n-s+1)} .$$

Hence

$$\begin{aligned}
\mathbf{Pr}\big(\text{rank}(\pi(x), \pi(X)) = r\big) &= \sum_{s=r}^{n-|X|+r} \frac{\binom{|X|-1}{r-1}\binom{n-|X|}{s-r}}{\binom{n}{s-1}(n-s+1)} \\
&= \frac{1}{|X|\binom{n}{|X|}} \sum_{s=r}^{n-|X|+r} \binom{s-1}{r-1}\binom{n-s}{|X|-r} \\
&= \frac{1}{|X|\binom{n}{|X|}} \binom{n}{|X|} = \frac{1}{|X|} .
\end{aligned}$$

(The second equality is obtained by expanding binomials into factorials and regrouping. The third equality is obtained by counting the ways of choosing $|X|$ elements out of $[n]$ by summing over all possible values s for the rth largest element among those chosen.) ∎

4. ROBUST FAMILIES

We now consider *robust* families. As described in the Introduction, robustness is an extension of min-wise independence. Formally, a family $\mathcal{P}$ is robust if for *every* possible permutation σ, when π is chosen at random in $\mathcal{P}$

$$\mathbf{Pr}\big(\min\{\sigma(\pi(X))\} = \sigma(\pi(x))\big) = \frac{1}{|X|} . \tag{10}$$

Trivially, S_n is a robust family. We first demonstrate that there exist non-trivial robust families. To this end, we extend the condition for min-wise independent families given in Lemma 5 to the equivalent condition for robust families. Since robust

families are min-wise independent under any order σ we obtain the following:

Lemma 6. *A family of permutations $\mathcal{P}$ is robust if and only if for any set $X \subset [n]$ of size k and any element $x \in [n] \setminus X$, and any other set $A \subset [n]$ of size also k and any element $a \in [n] \setminus A$*

$$\mathbf{Pr}\big(\pi(X) = A \wedge \pi(x) = a\big) = \frac{1}{\binom{n}{k}(n-k)} . \tag{11}$$

Theorem 3. *There exist biased robust families of size at most*

$$n^2\binom{2(n-1)}{n-1} .$$

Proof. Following an idea used in [5] (apparently used first in [11]), we establish a linear program for determining a robust family of the required size. There are $n!$ variables x_{π_i}, one for each possible permutation π_i. The variable x_{π_i} represents the probability that π_i is chosen within our family; if $x_{\pi_i} = 0$, we may exclude π_i from the family.

Our linear program is based on Lemma 6. We set up an equation for each pair (a, A) and (x, X) with $|A| = |X|$, with each equation representing the constraint that (a, A) maps to (x, X) with the required probability. Hence there are

$$\sum_{i=0}^{n-1} n^2\binom{n-1}{i}^2 = n^2\binom{2(n-1)}{n-1}$$

equations. We know there exists a solution to the linear program, since if each permutation is chosen with probability $1/n!$ we have a robust family. Hence there must be a basic feasible solution with at most $n^2\binom{2(n-1)}{n-1}$ variables taking nonzero values. This solution yields a biased robust family. ∎

It is also worthwhile to ask if there are any nontrivial unbiased robust families. We demonstrate that in fact there are nontrivial families for $n \geq 4$.

Recall that the permutations S_n can be split into two groups, each of size $n!/2$, as follows: A permutation is called *even* if it can be obtained by an even number of transpositions from the identity, and is called *odd* otherwise.

Theorem 4. *For $n \geq 4$, the even permutations and the odd permutations of $[n]$ both yield robust families.*

Proof. We use Lemma 6. That is, we must show that for each pair (x, X) with $x \in [n]$, $X \subseteq [n]$, $x \notin X$, the probability that $\pi(x) = a$ and $\pi(X) = A$ is correct for every (a, A) with $a \in [n]$, $A \subseteq [n]$, $|A| = |X|$, and $a \notin A$.

Equivalently, since the odd permutations and even permutations divide the set of all permutations into two equal-sized families, it suffices to show that the number of even permutations mapping (x, X) into (a, A) is the same as the number of odd permutations that do so. Note that as $n \geq 4$, either $|X| \geq 2$ or $|[n] - X - \{x\}| \geq 2$. In the first case, we can determine a one-to-one mapping of even permutations to odd permutations that map (x, X) into (a, A) by choosing two particular elements of X (say the two smallest) and transposing them. In the second case, we may do the same by transposing two elements of $[n] - X - \{x\}$. ∎

From the lower bound in [5], we know that unbiased min-wise independent families (and hence robust families) have size at least $\text{lcm}(1, \ldots, n)$. As $\text{lcm}(1, \ldots, n) = n!/2$ for $n = 4$ and $n = 5$, the result of Theorem 4 is optimal for these cases. We suspect that Theorem 4 is in fact optimal for all $n \geq 4$; that is, there is no unbiased robust family of size less than $n!/2$. While we cannot yet show this, we can show that for $n = 6$, there is no unbiased robust family of size $\text{lcm}(1, \ldots, n) = 60$.

Theorem 5. *All the unbiased robust families of permutations of* $\{1, 2, 3, 4, 5, 6\}$ *have size greater than* 60.

Proof. The smallest possible robust family has size 60, so we simply show that no such family of this size exists. The proof uses an exhaustive search, where the search is reduced using symmetry and Lemma 6.

Assume that an unbiased robust family of size 60 exists. Let us use the following shorthand: we write, for example, 342156 to represent the permutation π on $\{1, 2, 3, 4, 5, 6\}$ where $\pi(1) = 3$, $\pi(2) = 4$, etc. In this form, by Lemma 6 there must be 10 permutations in the family that begin with 1; in fact, by Lemma 6 there must be two permutations that begin with 12, two that begin with 13, etc. Now without loss of generality, we may assume (by symmetry) that two of the permutations in our family begin with 123 and 124.

Now, if some permutation in our family begins with 123, then no other permutation in our family can begin with 132, or Lemma 6 would be contradicted. Hence there are three possibilities for the two permutations in our family that begin with 13, namely 134, 135, and 136. Similarly, there are three possibilities for the two permutations in our family that begin with 14, namely 143, 145, and 146. Again by Lemma 6, both 134 and 143 cannot begin permutations in our family, so again without loss of generality we take 143, 135, and 136 to begin permutations in our family. Similarly we must then choose whether to take 145 or 146 in our family, and without loss of generality by symmetry we may take 145.

To this point, we have shown that, without loss of generality, we may assume our permutation family of size 60 has 10 permutations with the following prefixes:

123, 124, 135, 136, 143, 145, 152, 156, 162, 164.

We may then attempt to find valid completions to these 10 prefixes using an exhaustive computer search. For a set of completions to be valid, each permutation must in fact be a valid permutation of the numbers 1 through 6. Also, by Lemma 6, for the 10 permutations that begin with a 1, every other pair of positions must contain each unordered pair of the numbers 2 through 6 exactly once. These conditions make the exhaustive search process relatively simple. We find six possibilities, given in the top of Table 4.

To complete the proof, we now consider the 10 permutations with a 1 in the second position in our family. Since we have assumed that permutations with prefixes 123 and 124 are in our family, the two permutations that begin with 21 must be 215 and 216, by Lemma 6. Similarly, we find our permutation family must have 10 permutations with the following prefixes:

215, 216, 312, 314, 412, 416, 513, 514, 613, 615.

TABLE 1 Possible Subsets of a Robust Permutation Family.

123654	123465	123546	123456	123645	123564
124365	124536	124653	124563	124356	124635
135642	135264	135426	135246	135624	135462
136425	136542	136254	136524	136452	136245
143526	143652	143265	143625	143562	143256
145263	145326	145632	145362	145236	145623
152346	152634	152463	152643	152364	152436
156234	156423	156342	156432	156243	156324
162453	162345	162534	162354	162435	162543
164532	164253	164325	164235	164523	164352
215634	215463	215346	215436	215643	215364
216345	216534	216453	216543	216354	216435
312465	312546	312654	312564	312456	312645
314652	314265	314526	314256	314625	314562
412536	412653	412365	412635	412563	412356
416253	416325	416532	416352	416235	416523
513264	513426	513642	513462	513246	513624
514326	514632	514263	514623	514362	514236
613542	613254	613425	613245	613524	613452
615423	615342	615234	615324	615432	615243

Again we now complete these prefixes using exhaustive search; the six possibilities are given in the bottom of Table 4.

We now consider the 36 possible sets of 20 permutations obtained by taking one solution from the top and one solution from the bottom of Table 4. It is straightforward to check that in each of the 36 combinations Lemma 6 is violated. Hence our initial assumption that a robust familiy of permutations on {1, 2, 3, 4, 5, 6} of size 60 exists must be incorrect. ■

Given the development of approximate min-wise independent families of permutations developed in [5], it is natural to ask about approximate robust families of permutations as well. A family of permutations is said to be *approximately robust with relative error* ϵ if and only if for every permutation order σ,

$$\left|\mathbf{Pr}\big(\min\{\sigma(\pi(X))\} = \sigma(\pi(x))\big) - \frac{1}{|X|}\right| \leq \frac{\epsilon}{|X|}\ . \tag{12}$$

That is, regardless of σ, the probability over the choice of π that an element x is the minimum of a set $|X|$ is within a factor of $(1 \pm \epsilon)$ of the natural probability $1/|X|$. It is straightforward to show that there must be small approximate robust families.

Theorem 6. *There are approximate robust families of size* $O(n^2 \log(n)/\epsilon^2)$.

Proof. The proof follows Theorem 3 of [5]. We simply choose a random set of permutations of the appropriate size, and show that with some probability, we obtain an unbiased approximate robust family.

For a permutation π chosen uniformly at random from S_n,

$$\mathbf{Pr}(\sigma(\pi(x)) = \min \sigma(\pi(X))) = \frac{1}{|X|}.$$

Suppose we pick f permutations uniformly at random from S_n. Consider a permutation σ, a set X, and an element $x \in X$. Let $A(\sigma, x, X)$ be the number of permutations for which $\sigma(\pi(x)) = \min \sigma(\pi(X))$. Note that $A(\sigma, x, X)$ has the binomial distribution Bin$(f, 1/|X|)$. Then $E[A(\sigma, x, X)] = f/|X|$. Let $B(\sigma, x, X)$ be the event $|A(\sigma, x, X) - (f/|X|)| > \epsilon(f/|X|)$. The event $B(\sigma, x, X)$ is considered a *bad* event for the triple (σ, x, X). We will be interested in bounding the probability of bad events. Applying Chernoff bounds (see for example [1]), we have

$$\mathbf{Pr}(B(\sigma, x, X)) < 2e - (f_\epsilon^2/3|X|) \leq 2e - (f_\epsilon^2/3n).$$

This must hold for all triples (σ, x, X) such that $x \in X \subseteq [n]$. There are $n2^{n-1}n!$ such triples. Hence the probability that at least one bad event $B(\sigma, x, X)$ occurs is at most $n2^n n!e - (f\epsilon^2/3n$. For $f > 3n(\ln n! + n \ln 2 + \ln n)/\epsilon^2$, this probability is less than 1. Hence, for f this large, with nonzero probability no bad event occurs, and therefore there is some family of permutations that is approximately robust with relative error ϵ. ■

5. CONCLUSIONS

Our work raises several open questions. A more complete understanding of robust permutation families as well as families that are approximately robust or have approximate rank uniformity would be interesting, Another important question is whether our main result can be extended to approximate min-wise independent permutation families. Recall that we have shown that any sampling scheme (denoted by s) for the relative intersection of sets with the property

$$\mathbf{Pr}\big(s(A) = s(B)\big) = \frac{|A \cap B|}{|A \cup B|}$$

is equivalent to a min-wise independent family of permutations in Theorem 1. Can a similar relationship be shown with approximately min-wise independent families in the case where the sampling scheme is only approximate? That is, if we have a sampling scheme

$$(1 - \epsilon)\frac{|A \cap B|}{|A \cup B|} \leq \mathbf{Pr}\big(s(A) = s(B)\big) \leq (1 + \epsilon)\frac{|A \cap B|}{|A \cup B|},$$

must the sampling scheme naturally map to an approximately min-wise independent family?

ACKNOWLEDGMENT

We wish to thank Uri Feige for helpful suggestions.

REFERENCES

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

[2] A. Z. Broder, On the resemblance and containment of documents, Proc Compression and Complexity of Sequences 1997, 29, IEEE Computer Society, 1988, pp. 21–29.

[3] A. Z. Broder, Filtering near-duplicate documents. Preliminary version presented at FUN 98. Final version in Proc of the 11th Annual Symposium on Combinatorial Pattern Matching. June 2000. Available as Lecture Notes in Computer Science, Vol. 1848.

[4] A. Z. Broder, M. Burrows, and M. S. Manasse, Efficient computation of minima of random functions. Manuscript.

[5] A. Z. Broder, M. Charikar, A. Frieze, and M. Mitzenmacher, Min-wise independent permutations, Journal of Computer and System Sciences, 60 (2000), 630–659.

[6] A. Z. Broder, M. Charikar, and M. Mitzenmacher, A derandomization using min-wise independent permutations, Proc Random 98, 1998, pp. 15–24. available as Lecture Notes in Computer Science, Vol. 1518.

[7] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, Syntactic clustering of the Web, Proc Sixth International World Wide Web Conference, 1997, pp. 391–404.

[8] E. Cohen, Size-estimation framework with applications to transitive closure and reachability. Journal of Computer and System Sciences, 55 (1997), 441–453.

[9] P. Indyk, A small approximately min-wise independent family, Proc 10th ACM-SIAM Symp Discrete Algorithms, 1999, pp. 454–456.

[10] D. Karger, E. Lehman, T. Leighton, M. Levine, D. Lewin, and R. Panigrahy, Consistent hashing and random trees: Distributed caching protocols for relieving hot spots on the World Wide Web, Proc 29th ACM Symp Theory of Computing, El Paso, Texas, 1997, pp. 654–663.

[11] D. Koller and N. Megiddo, Constructing small sample spaces satisfying given constraints, SIAM J Discrete Math, 7 (1994), pp. 260–274.

[12] K. Mulmuley, Randomized geometric algorithms and pseudorandom generators, Algorithmica, 16 (1996), 450–463.

[13] M. Saks, A. Srinivasan, S. Zhou, and D. Zuckerman, Low discrepancy sets yield approximate min-wise independent permutation families, Information Processing Letters, 73:1–2 (2000), pp. 29–32.

[14] Y. Takei, T. Itoh, and T. Shinozaki, An optimal construction of exactly min-wise independent permutations, Technical Report COMP98-62, IEICE, 1998.