Min-Wise Independent Permutations

Andrei Z. Broder[*] Moses Charikar[†] Alan M. Frieze[‡]
Michael Mitzenmacher[§]

Abstract

We define and study the notion of min-wise independent families of permutations. We say that $\mathcal{F} \subseteq S_n$ is *min-wise independent* if for any set $X \subseteq [n]$ and any $x \in X$, when π is chosen at random in $\mathcal{F}$ we have

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

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.

Our research was 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. However, in the course of our investigation we have discovered interesting and challenging theoretical questions related to this concept – we present the solutions to some of them and we list the rest as open problems.

[*] Digital SRC, 130 Lytton Avenue, Palo Alto, CA 94301, USA. E-mail: broder@pa.dec.com.

[†] Computer Science Department, Stanford University, CA 94305, USA. E-mail: moses@cs.stanford.edu. Part of this work was done while this author was a summer intern at Digital SRC. Supported by the Pierre and Christine Lamond Fellowship and in part by an ARO MURI Grant DAAH04-96-1-0007 and NSF Award CCR-9357849, with matching funds from IBM, Schlumberger Foundation, Shell Foundation, and Xerox Corporation.

[‡] Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA. Part of this work was done while this author was visiting Digital SRC. Supported in part by NSF grant CCR9530974. E-mail: af1p@andrew.cmu.edu

[§] Digital SRC, 130 Lytton Avenue, Palo Alto, CA 94301, USA. E-mail: michaelm@pa.dec.com.

1 Introduction

The classic analysis of hashing schemes often entails the assumption that the hash functions used are random. More precisely, the assumption is that keys belonging to a universe $\mathcal{U}$ are hashed into a table of size M by choosing a function h uniformly at random among all the functions $\mathcal{U} \to [M]$. (The notation $[M]$ stands for the set $\{0, \ldots, M-1\}$. This is slightly non-standard, but convenient for our purposes.) This assumption is impractical since just specifying such a function requires $|\mathcal{U}| \log(M)$ bits[1], which usually far exceeds the available storage.

Fortunately in most cases heuristic hash functions behave very closely to the expected behavior of random hash functions; but there are cases when rigorous probabilistic guarantees are necessary. For instance, various adaptive hashing schemes presume that a hash function with certain prescribed properties can be found in constant expected time. This holds if the function is chosen uniformly at random from all possible functions until a suitable one is found but not necessarily if the search is limited to a smaller set of functions.

This situation has led Carter and Wegman [13] to the concept of *universal hashing*. A family of hash functions $\mathcal{H}$ is called *weakly universal* if for any pair of distinct elements $x_1, x_2 \in \mathcal{U}$, if h is chosen uniformly at random from $\mathcal{H}$ then

$$\mathbf{Pr}(h(x_1) = h(x_2)) \leq \frac{1}{|M|} \tag{1}$$

and is called *(strongly) universal* or *pair-wise independent* if for any pair of distinct elements $x_1, x_2 \in \mathcal{U}$ and arbitrary $y_1, y_2 \in [M]$

$$\mathbf{Pr}(h(x_1) = y_1 \text{ and } h(x_2) = y_2) = \frac{1}{|M|^2}. \tag{2}$$

It turns out that in many situations the analysis of various hashing schemes can be completed under the weaker assumption that h is chosen uniformly at random from a universal family, rather than the assumption that h is chosen uniformly at random from among all possible functions. In other words, limited randomness suffices. Furthermore, there exist universal families of size $O(|M|^2)$ that can be easily implemented in practice. Thus, universal hash functions are very useful in the design of adaptive hash schemes (see e.g. [12, 16]) and are actually used in commercial high-performance products (see e.g. [24]). Moreover, the concept of

[1] We use log for $\log_2$ throughout.

pairwise independence has important theoretical applications. (See the excellent survey by Luby and Wigderson [22].)

It is often convenient to consider permutations rather than functions. Let S_n be the set of all permutations of $[n]$. We say that a family of permutations $\mathcal{F} \subseteq S_n$ is *pair-wise independent* if for any $\{x_1, x_2, y_1, y_2\} \subseteq [n]$ with $x_1 \neq x_2$ and $y_1 \neq y_2$,

$$\mathbf{Pr}(\pi(x_1) = y_1 \text{ and } \pi(x_2) = y_2) = \frac{1}{n(n-1)}. \tag{3}$$

In a similar vein, in this paper, we say that $\mathcal{F} \subseteq S_n$ is *exactly min-wise independent* (or just *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 $\mathcal{F}$ we have

$$\mathbf{Pr}(\min\{\pi(X)\} = \pi(x)) = \frac{1}{|X|}. \tag{4}$$

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. Unless otherwise stated we shall assume that π is chosen uniformly at random in $\mathcal{F}$; otherwise, we shall say π is chosen with a *biased* distribution μ. Uniform distributions are natural in this setting, since in practice they are simple to represent.

As explained below, this definition is motivated by the fact that such a family (under some relaxations) is essential to the algorithm currently used in practice by the AltaVista Web indexing software [23] to detect and filter near-duplicate documents.

The Web [5] has undergone exponential growth since its birth, and this has lead to the proliferation of documents that are identical or near identical. Experiments indicate that over 20% of the publicly available documents on the web are duplicates or near-duplicates. These documents arise innocently (e.g. local copies of popular documents, mirroring), maliciously (e.g., “spammers” and “robot traps”), and erroneously (spider mistakes). In any case they represent a serious problem for indexing software for two main reasons: first, indexing of duplicates wastes expensive resources; and second, users are seldom interested in seeing documents that are “roughly the same” in response to their queries.

This informal concept does not seem to be well captured by any of the standard distances defined on strings (Hamming, Levenshtein, etc.). Furthermore the computation of these distances usually requires the pairwise comparison of entire documents. For a very large collection of documents this is not feasible, and a sampling mechanism per document is necessary.

It turns out that the problem can be reduced to a set intersection problem by a process called *shingling*. (See [7, 11] for details.) Via shingling each document D gets an associated set S_D. For the purpose of the discussion here we can view S_D as a set of natural numbers. (The size of S_D is about equal to the number of words in D.) The *resemblance* $r(A, B)$ of two documents, A and B, is defined as

$$r(A, B) = \frac{|S_A \cap S_B|}{|S_A \cup S_B|}.$$

Experiments seem to indicate that high resemblance (that is, close to 1) captures well the informal notion of "near-duplicate" or "roughly the same".

To compute the resemblance of two documents it suffices to keep for each document a relatively small, fixed size *sketch*. The sketches can be computed fairly fast (linear in the size of the documents) and given two sketches the resemblance of the corresponding documents can be computed in linear time in the size of the sketches.

This is done as follows. Assume that for all documents of interest $S_D \subseteq \{1, \ldots, n\}$. (In practice $n = 2^{64}$.) Let π be chosen uniformly at random over S_n, the set of permutations of $[n]$. Then

$$\mathbf{Pr}(\min\{\pi(S_A)\} = \min\{\pi(S_B)\}) = \frac{|S_A \cap S_B|}{|S_A \cup S_B|} = r(A, B). \tag{5}$$

Hence, we can choose say, 100 independent random permutations $\pi_1, \ldots, \pi_{100}$. For each document D, we store the list

$$\bar{S}_A = (\min\{\pi_1(S_A)\}, \min\{\pi_2(S_A)\}, \ldots, \min\{\pi_{100}(S_A)\}).$$

Then we can readily estimate the resemblance of A and B by computing how many corresponding elements in $\bar{S}_A$ and $\bar{S}_B$ are common. (For a set of documents, we avoid quadratic processing time, because a particular value for any coordinate is usually shared by only a few documents. For details see [7, 8, 11].)

In practice, as in the case of hashing discussed above, we have to deal with the sad reality that it is impossible to choose π uniformly at random in S_n. We are thus led to consider smaller families of permutations that still satisfy the min-wise independence condition given by equation (4), since min-wise independence is necessary and sufficient for equation (5) to hold.

In practice we can allow certain relaxations. First, we can accept small relative errors. We say that $\mathcal{F} \subseteq S_n$ 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 $\mathcal{F}$ we have

$$\left|\mathbf{Pr}(\min\{\pi(X)\} = \pi(x)) - \frac{1}{|X|}\right| \leq \frac{\epsilon}{|X|}. \tag{6}$$

In other words we require that all the elements of any fixed set X have only an almost equal chance to become the minimum element of the image of X under π. The expected relative error made in evaluating resemblance using approximately min-wise independent families is less than ϵ.

Second, the sets of interest are usually much smaller than n. (For the situation discussed above the typical set has size 1000 while $n = 2^{64}$.) We say that $\mathcal{F} \subseteq S_n$ is *restricted min-wise independent for sets up to size k* (or just restricted min-wise independent where the meaning is clear) if for any set $X \subseteq [n]$ with $|X| \leq k$ and any $x \in X$, when π is chosen at random in $\mathcal{F}$ we have

$$\mathbf{Pr}(\min\{\pi(X)\} = \pi(x)) = \frac{1}{|X|}, \qquad |X| \leq k. \tag{7}$$

Of course we can consider families that are both restricted and approximately min-wise independent.

Third and finally, it turns out that whether the distribution on the family $\mathcal{F}$ is uniform or not leads to qualitatively different results.

Ultimately we are interested in practical families of permutations. Hence we first study what is the minimum size of a family that satisfies various combinations of requirements. Clearly if the minimum size is exponential no practical solution exists. It turns out that the exact min-wise property generally necessitates exponential size but that the approximate property can be satisfied by polynomial size families. The complete synopsis of our results is given in Table 1. The entries for which we have no bounds beyond those implied by other entries in the table are marked "?" and the entries for which we have no non-trivial bounds are marked "???".

Starting from the opposite end we study how good is the performance provided by various families that are easily implementable in software. We consider pair-wise independent families, for which there are numerous practical implementations. In particular we are interested in linear transformations, since they are used in the AltaVista implementation and are known to perform better in some situations than other pair-wise independent families (see [1]).

The way we evaluate this performance is to consider a set X and study the distribution of the minimum of the image of X. It suffices to examine the two elements that are respectively most likely and least likely to become the minimum

Family type	**Upper bound**	**Lower bound**
Exactly min-wise, uniform distrib on $\mathcal{F}$	4^n	$e^{n-o(n)}$
Exactly min-wise, biased distrib on $\mathcal{F}$	$n2^{n-1} - 1$	$\Omega\left(\sqrt{n}\, 2^n\right)$
Exactly min-wise, restricted, uniform distrib on $\mathcal{F}$	?	$e^{k-o(k)}$
Exactly min-wise, restricted, biased distrib on $\mathcal{F}$	$\sum_{j \le k} j \binom{n}{j}$	$\Omega\left(k2^{k/2}\log\left(\frac{n}{k}\right)\right)$
Approx min-wise, uniform distrib on $\mathcal{F}$	$O\left(n^2/\epsilon^2\right)$ (existential) ??? (constructive)	$n^2\left(1-\sqrt{8\epsilon}\right)$
Approx min-wise, biased distrib on $\mathcal{F}$	???	$\max_{r \ge 1} \frac{(n-r)\binom{n}{r}}{1+\epsilon\binom{n}{r}}$
Approx min-wise, restricted, uniform distrib on $\mathcal{F}$	$O\left(\frac{k^2\log(n/k)}{\epsilon^2}\right)$ (existential) $2^{4k+o(k)}k^{2\log(\log n/\epsilon)}$ (constructive)	?
Approx min-wise, restricted, biased distrib on $\mathcal{F}$	?	$\Omega\left(\min\left(k2^{k/2}\log(n/k), \frac{\log(1/\epsilon)\left(\log n - \log\log(1/\epsilon)\right)}{\epsilon^{1/3}}\right)\right)$

Table 1: Synopsis of results – minimum size of families

since all the other elements will become the minimum with a probability in between the extremal values. We consider two situations: when X is chosen to be the worst set (farthest from uniform) with regard to the property of interest; and when X is chosen uniformly at random, in which case we look for the expected value of the bound over the random choices of X. The synopsis of our answers is given in Table 2, where we follow the same convention as before regarding the use of "?" and "???".

Finally, we note that while our definition of min-wise independence and our

Family type	**Bounds on the most probable element**		**Bounds on the least probable element**	
	Upper	**Lower**	**Upper**	**Lower**
Pairwise independent – worst set	$O\left(\frac{1}{\sqrt{k}}\right)$	?	???	$\frac{1}{2(k-1)}$
Linear – worst set	?	$\frac{3}{\pi^2}\frac{\ln k}{k}$	$\frac{12\ln 2}{\pi^2\, k}$	?
Pairwise independent – random set	$\frac{1+1/\sqrt{2}}{k}$	???	???	?
Linear – random set	?	???	???	?

Table 2: Synopsis of results – quality of approximation

subsequent results appear novel, similar ideas have appeared in the literature. For example, the property of min-wise independence appears to be a key feature of the monotone ranged hash functions described in [19]. Cohen uses the properties of the minimum element of a random permutation to estimate the size of the transitive closure, as well as to solve similar related problems [14]. Given these connections, as well as the history of the development of pairwise independence, we expect that the concept of min-wise independence will prove useful in many future applications.

A preliminary version of this work has appeared in [9]. Since then new constructions have been proposed by Indyk [18] and others [25]. The use of min-wise independent families for derandomization is discussed in [10].

2 Exact Min-Wise Independence

In this section, we provide bounds for the size of families that are exactly min-wise independent. We begin by determining a lower bound, demonstrating that the size of the family $\mathcal{F}$ must grow exponentially with n.

Theorem 1 *Let $\mathcal{F}$ be min-wise independent. Then $|\mathcal{F}|$ is at least as large as the least common multiple (lcm) of the numbers $1, 2, \ldots n$, and hence $|\mathcal{F}| \geq e^{n-o(n)}$.*

Proof: Let X be a subset of $[n]$ with $|X| = j$. Each element of X must be the minimum under the family $\mathcal{F}$ the same number of times, so j must divide $|\mathcal{F}|$. This holds for every $j \in \{1, 2, \ldots n\}$, so the lcm of $\{1, 2, \ldots n\}$ must divide $|\mathcal{F}|$. That the lcm of the first n numbers has size $e^{n-o(n)}$ is a well known fact of number theory [4, p. 76]. $\square$

Remark 1 *This proof also gives a lower bound of $e^{k-o(k)}$ for restricted min-wise independent families. Also, note that the proof does not require that the members of $\mathcal{F}$ be distinct. Hence the theorem holds even if $\mathcal{F}$ contains duplicates of some permutations.*

We now describe a min-wise independent family of size less than 4^n, which is significantly smaller than the trivial bound of $n!$ and of the same form as the lower bound given above.

Theorem 2 *There exists a min-wise independent family $\mathcal{F}$ of size less than 4^n.*

Proof: We initially assume for convenience, that $n = 2^r$ for some r. We construct the family of permutations recursively in stages. In the first stage, we divide the set $[n]$ into two equal halves, the top and the bottom. At the first stage, there are $\binom{n}{n/2}$ ways to partition the set. Each of these can be described by an n bit string with exactly $n/2$ ones in it. Element i goes in the top half if and only if the bit string has a 1 in the ith position. We proceed to partition each half. Again this can be done by choosing a $n/2$ bit string with $n/4$ ones in it. There are $\binom{n/2}{n/4}$ such strings. Importantly, we use the same string for each half. At the ith stage, we have the set divided into 2^{i-1} parts each of size $n/2^{i-1}$. We partition each part into two halves by choosing a $n/2^{i-1}$ bit string with $n/2^i$ ones and using this string to define the partition for each of the 2^{i-1} parts. We continue in this way until each part has size 1. This process produces a permutation of the set in a natural way, with the topmost element receiving the smallest number in the permutation.

The property that each element is the minimum with the correct probability can be verified directly by calculation. More intuitively, when we split $[n]$ into two halves, every element of X has an equal chance to go to the upper half or to the lower half; furthermore, all elements of X now in the top half are equally likely to eventually become the topmost element of X (by induction). If no elements of X are in the top half, then all lie in the bottom, and again (by induction) all are equally likely to become eventually the topmost.

The number of permutations in this family is

$$\prod_{i=1}^{\log n} \binom{n/2^{i-1}}{n/2^i}.$$

A simple calculation shows that the size of this family is $4^{n-O(\log^2 n)}$.

We now explain how to remove the assumption that n is a power of 2. Earlier, we used the fact that a j bit string with $j/2$ ones defines a partition of a set of size j into two equal halves. We now use the that fact a j bit string with $l \geq j/2$ ones defines a partition of any set of size $r \leq j$ into two parts such that each is of size at most l. We construct the permutations in stages as before. At the beginning of the ith stage, we have partitioned the set into 2^{i-1} parts, each of size at most $\lceil \frac{n}{2^{i-1}} \rceil$. We continue by choosing a string of length $\lceil \frac{n}{2^{i-1}} \rceil$ with $\lceil \frac{n}{2^i} \rceil$ ones. We use this to partition each of the 2^{i-1} parts into two, such that the maximum size of the parts produced is at most $\lceil \frac{n}{2^i} \rceil$. We perform this partition for $\lceil \log n \rceil$ stages, giving us a min-wise independent permutation of $[n]$. The number of possible permutations is[2]

$$\prod_{i=1}^{\lceil \log n \rceil} \binom{\lceil \frac{n}{2^{i-1}} \rceil}{\lceil \frac{n}{2^i} \rceil},$$

and hence the size of this family is also less than 4^n. □

Remark 2 *It is worth noting that this family has much stronger properties than what we actually require. For example, if instead of looking at just the minimum element, we look at the unordered set of the smallest j elements for any $j \leq |X|$, this unordered set is equally likely to be any subset of X of size j.*

2.1 Exact problem with non-uniform distribution

Although we focus on results for uniform distributions, we demonstrate here an interesting result: the lower bound of Theorem 1 can be beaten by using non-uniform distributions.

Theorem 3 *There is a family $\mathcal{F}$ of size at most $n2^{n-1} - 1$, such that $\mathcal{F}$ with an associated distribution μ is min-wise independent.*

[2] Proving directly that this number is a multiple of $\text{lcm}(1, \ldots, n)$ is an amusing exercise, at least for certain people.

Proof: We can write a linear program to find a $\mathcal{F}$ and μ satisfying the theorem. We have a variable x_π for each of the permutations $\pi \in S_n$, where x_π represents the weight of π according to μ. For every $X \subset [n]$ and for every $x \in X$, we express the condition that $\mathbf{Pr}(\min\{\pi(X)\} = \pi(x)) = \frac{1}{|X|}$ as a linear equation in the variables x_π. We have a total of $\sum_{k=1}^{n} k \cdot \binom{n}{k} = n2^{n-1} - 1$ constraints. This system clearly has a feasible solution (choose an element of S_n uniformly at random; that is, put $x_\pi = 1/n!$ for all $\pi \in S_n$), and hence it has a basic feasible solution with at most $n \cdot 2^{n-1} - 1$ non-zero variables. This solution yields a family satisfying the conditions of the theorem. □

Remark 3 *Although Theorem 3 beats the lower bound of Corollary 1, the size of the family is still exponential in n, and we will prove an almost tight lower bound in Section 3.4. Also, for restricted min-wise independence, this same construction gives an upper bound of* $\sum_{j=1}^{k} j \cdot \binom{n}{j}$.

3 The Approximate Problem

As the exact problem requires exponential sized families, we turn our attention to the approximate problem. In this section, we prove some existential upper bounds and constructive upper bounds as well as derive lower bounds for the approximate problem.

3.1 Existential Upper Bounds

We obtain existential upper bounds on the sizes of approximately min-wise independent families via the probabilistic method [3], by simply choosing a number of random permutations from S_n.

Theorem 4 *There exist families of size* $O(\frac{n^2}{\epsilon^2})$ *that are approximately min-wise independent and there exist families of size* $O(\frac{k^2 \ln(n/k)}{\epsilon^2})$ *that are approximately and restricted min-wise independent.*

Proof: Assume $0 \leq \epsilon \leq 1$. We apply a straightforward probabilistic argument. Suppose we pick f permutations uniformly at random from S_n. Consider a set X and an element $x \in X$. For a permutation π chosen uniformly at random, $\mathbf{Pr}(\pi(x) = \min \pi(X)) = \frac{1}{|X|}$. Let $A(x, X)$ be the number of permutations for which $\pi(x) = \min \pi(X)$. Note that $A(x, X)$ has the binomial distribution $Bin(f, \frac{1}{|X|})$.

Then $E[A(x,X)] = \frac{f}{|X|}$. Let $B(x,X)$ be the event $|A(x,X) - \frac{f}{|X|}| > \epsilon\frac{f}{|X|}$. The event $B(x,X)$ is considered a *bad* event for the pair (x,X). We will be interested in bounding the probability of bad events. Applying Chernoff bounds (see for example [3]), we have

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

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

For the restricted case where $|X| \leq k$, the same argument holds, except now

$$\mathbf{Pr}(B(x,X)) < 2e^{-\frac{f\epsilon^2}{3|X|}} \leq 2e^{-\frac{f\epsilon^2}{3k}}.$$

Again his must hold for all suitable pairs (x,X), but as $|X| \leq k$, there are only $\sum_{i=1}^{k} i \cdot \binom{n}{i} < \sum_{i=0}^{k} i \cdot \binom{n+i}{i} = \binom{n+k+1}{k}$ such pairs. Hence the probability that at least one bad event $B(x,X)$ occurs is at most $2\binom{n+k+1}{k}e^{-\frac{f\epsilon^2}{3k}}$. For

$$f > \frac{3k}{\epsilon^2}\ln\binom{n+k+1}{k} + \ln 2,$$

this probability is less than 1, and this implies the second part of the theorem. □

Remark 4 *Of course the above argument can also be used to show that selecting $O(\frac{n^2}{\epsilon^2})$ permutations uniformly at random yields an approximately min-wise independent family with high probability. Moreover, the permutations need not be chosen uniformly at random from S_n, but could instead be chosen from any family that yields exact min-wise independence, such as the family given in Theorem 2. Although this would appear to provide a suitable solution for the document similarity problem discussed in the introduction, in practice, this result does not help us. The problem is that one cannot conveniently represent a random permutation from S_n. Recall that a random permutation on n elements requires on average $\Omega(n \log n)$ bits to represent, and in practice $n = 2^{64}$. This leads us to consider simple linear permutations in Section 4.*

3.2 Constructive Upper Bounds

Although the techniques of the last section show that sufficiently large families chosen at random will be approximately and restricted min-wise independent with high probability, they do not appear to provide a way to explicitly construct a suitable family. In fact, we do not even know of an efficient procedure to check that a randomly chosen family is approximately and restricted min-wise independent for given families of k and ϵ. Hence here we provide an explicit construction.

Theorem 5 *There exists an approximately and restricted min-wise independent family $\mathcal{F}$ of size $2^{4k+o(k)}k^{2\log\frac{\log n}{\epsilon}}$.*

Proof: The idea, similar to that in Theorem 2, is to split the set up into groups. Instead of initially splitting the set $[n]$ into two equal groups, however, we instead split the set $[n]$ into r random groups for a suitable r using a k-wise independent hash function. Since we are concerned with sets X of size at most k, it is likely that a k-wise independent hash function will divide the elements of X so that no more than $k/2$ fall in any hash bucket. We then continue recursively.

Our construction of these hash functions is based on the explicit construction of almost k-wise independent distributions on N bit binary strings. We use the following result from [2]:

Proposition 1 *We can construct a family of N bit strings which are δ away (in the L_1 norm) from k-wise independence, such that $\log|\mathcal{F}|$ is at most $k+2\log(\frac{k\log N}{2\delta})+2$.*

We use this proposition to construct an almost k-wise independent family of hash functions from $[n]$ to $[r]$, where we choose a suitable value of r later. A hash function mapping $[n]$ to $[r]$ can be described by a string of length $N = n\log r$ bits, using $\log r$ bits to determine the image of each of the n elements in the domain. Further, if the family of N bit strings is $k\log r$-wise independent, the family of hash functions is k-wise independent. Each hash function h defines a permutation $\sigma_h \in S_n$ as follows: for a hash function h, we sort all the elements of $[n]$ in the order $(h(x), x)$, i.e. x_1 occurs before x_2 if either $h(x_1) < h(x_2)$ or $h(x_1) = h(x_2)$ and $x_1 < x_2$. The sorted order defines the permutation σ_h.

Suppose temporarily that our family of hash functions were exactly k-wise independent. Fix a set X of size k. We consider a hash function to be good if all the elements of X are hashed to distinct locations, and bad otherwise. Since the family of hash functions is k-wise independent, for any two elements x_1, x_2 of X, the probability that $h(x_1) = h(x_2)$ is $1/r$. The probability that two elements of

X hash to the same location is thus at most $\frac{k^2}{2r}$, and therefore the fraction of bad hash functions is at most $\frac{k^2}{2r}$. Thus, for the family of permutations obtained, the probability of any element being the minimum deviates from the mean by at most $\frac{k^2}{2r}$. If the bit strings used to construct the hash functions are actually δ away (in the L_1 norm) from being $k \log r$-wise independent, the deviation from the mean is at most $\delta + \frac{k^2}{2r}$. Choosing $\delta = \frac{\epsilon}{2k}$ and $r = \frac{k^3}{\epsilon}$ yields a deviation at most $\frac{\epsilon}{k}$ as desired.

We obtain a smaller family by breaking the process of hashing $[n]$ to $[r]$ into several steps, again in the spirit of Theorem 2. We use t hash functions h_i, $1 \leq i \leq t$, such that h_i hashes $[n]$ to $[r_i]$, and now $r = \prod_{i=1}^{t} r_i$. We can view h_1 as selecting the most significant bits of the hash value of each element, h_2 as selecting the next most significant bits, and so on. Although we need h_1 to be k-wise independent, we can use less independence with each successive h_i, yielding a smaller family.

For our construction, we will choose h_i to be almost k_i-wise independent, where $k_1 = k$ and $k_{i+1} = k_i/2$. We choose r_i so that h_i maps any set of size k_i into $[r_i]$ in such a way so that no bucket has size greater than k_{i+1} with probability at least $1 - \frac{\epsilon}{2\lceil \log k \rceil}$. We choose the h_i close enough to k_i-wise independent so that the difference adds an error probability $\frac{\epsilon}{2\lceil \log k \rceil}$ per level. For convenience we replace $\lceil \log k \rceil$ by $\log k$ in the derivation below; the difference is absorbed in the order notation.

If h_i were exactly k_i-wise independent, the probability of having more than $k_i/2$ elements hashed to any location would be

$$\sum_{l=k_i/2+1}^{k_i} \binom{k_i}{l} \left(\frac{1}{r_i}\right)^{l} \left(1 - \frac{1}{r_i}\right)^{k_i - l} \leq 2^{k_i} \left(\frac{1}{r_i}\right)^{k_i/2}.$$

For this to be less than $\frac{\epsilon}{2 \log k}$ requires

$$\log r_i \geq 2 + \frac{2}{k_i}\left(\log \frac{2}{\epsilon} + \log \log k\right)$$

or

$$k_i \log r_i \geq 2k_i + 2\left(\log \frac{2}{\epsilon} + \log \log k\right)$$

Hence, to generate h_i, we need an almost $k_i \log r_i$- wise independent distribution on $n \log r_i$ bits, where the distribution should be $\frac{\epsilon}{2 \log k}$ close to independent. From Proposition 1, this requires

$$b_i = k_i \log r_i + 2 \log \left(\frac{k_i \log r_i \log(n \log r_i) \log k}{\epsilon}\right) + 2 \text{ bits.}$$

Summing and ignoring lower order terms, we need $4k + 2(\log k)\log(\frac{\log k \log n}{\epsilon})$ total bits, yielding a suitable constructible family of size $2^{4k+o(k)}k^{2\log(\frac{\log n}{\epsilon})}$. □

3.3 Lower Bound for Uniform Families

We will prove a lower bound of $n^2(1-\sqrt{8\epsilon})$ for families with the uniform probability distribution. This shows that the n^2 term in the existential upper bound of Theorem 4 cannot be improved.

Theorem 6 *Let $\mathcal{F}$ be an approximate min-wise independent family. Then $|\mathcal{F}| \geq n^2(1-\sqrt{8\epsilon})$.*

Proof: Let $|\mathcal{F}| = f$. There must be some element a such that $\pi(a) = 1$ (that is, a is the second smallest after the permutation) for at least f/n permutations of $\mathcal{F}$. Fix such an a and consider $z \leq f/n$ such permutations. We will choose a value for z later. Let Z be the set of elements which occur as the smallest element in these z permutations (that is, $b \in Z$ iff $\pi(b) = 0$ for at least one of these z permutations) and let $S = [n] - Z$. Clearly $a \in S$ and $|S| \geq n - z$. Consider for how many permutations $\pi \in \mathcal{F}$ it is the case that $\pi(a)$ is the smallest element of $\pi(S)$. This happens at least whenever $\pi(a) = 0$ and also for the z permutations discussed above, where $\pi(a) = 1$ but an element not in S has image 0 under π. But $\pi(a) = 0$ for at least $\frac{f}{n}(1-\epsilon)$ permutations, because $\mathcal{F}$ is an approximately min-wise independent family; and for the same reason, $\pi(a)$ can be the minimum element of S for at most $\frac{f}{|S|}(1+\epsilon) \leq \frac{f(1+\epsilon)}{n-z}$ permutations. Hence

$$\frac{f(1-\epsilon)}{n} + z \leq \frac{f(1+\epsilon)}{n-z}.$$

Solving this equation for f and (almost) optimizing for z ($z = \sqrt{2\epsilon}f/n$) yields

$$f \geq n^2 \frac{1-\sqrt{2\epsilon}}{1+\sqrt{2\epsilon}-\epsilon}.$$

Simplifying the above yields a lower bound of $n^2(1-\sqrt{8\epsilon})$ on $|\mathcal{F}|$. □

3.4 Lower Bound for Non-Uniform Families

We will prove a lower bound on the size of any approximately min-wise independent family, even non-uniform families with an associated probability distribution μ. Our lower bound proof also yields a lower bound for non-uniform exactly min-wise independent families that is very close to the upper bound of $n2^{n-1} - 1$ obtained in Section 2.1.

Theorem 7 *Let $\mathcal{F}$ be an approximate min-wise independent family, possibly with an associated probability distribution μ. Then $|\mathcal{F}| \geq \frac{(n-r)\binom{n}{r}}{1+\epsilon 2^r \binom{n}{r}}$, for any $r < n$.*

Proof: Fix an element a and a set $Z = \{x_1, x_2, \ldots x_r\} \subseteq [n]$ with $a \notin Z$. Let us say that the pair (Z, a) is *satisfied* if there is a permutation π in $\mathcal{F}$ that has all the elements of $\pi(Z)$ as the r smallest elements of π in any order (that is, $\pi(Z) = [r]$) and has a as the $(r+1)$st smallest element (that is, $\pi(a) = r + 1$). We will show that most pairs (Z, a) must be satisfied for $\mathcal{F}$ to be an approximately min-wise independent family, and that in fact all pairs (Z, a) must be satisfied for $\mathcal{F}$ to be an exactly min-wise independent family,

Let $Y = [n] - Z$. By definition $a \in Y$. We consider the sets $Y_i = Y \cup x_i$ and count how often $\pi(a)$ is the smallest element of $\pi(Y_i)$. Let B_S be the event that a is the minimum of $\pi(S)$ when we choose a permutation from $\mathcal{F}$ under the distribution μ. Let $B = \bigcup_{i=1}^{r} B_{Y_i}$. Then $B \subseteq B_Y$, and hence $\mathbf{Pr}(B_Y - B) = \mathbf{Pr}(B_Y) - \mathbf{Pr}(B)$. On the other hand, the event $B_Y - B$ is precisely the event that (Z, a) is satisfied.

We now use the inclusion-exclusion principle to calculate $\mathbf{Pr}(B) = \mathbf{Pr}(\bigcup_{i=1}^{r} B_{Y_i})$. It is helpful to note the following facts. First if $a \in S_2 \subseteq S_1$ then $B_{S_1} \subseteq B_{S_2}$ and if $a \in S_1 \cap S_2$ then $B_{S_1} \cap B_{S_2} = B_{S_1 \cup S_2}$. Second, by the definition of approximate min-wise independence, $\frac{1-\epsilon}{|S|} \leq \mathbf{Pr}(B_S) \leq \frac{1+\epsilon}{|S|}$. We will abbreviate this by saying that $\mathbf{Pr}(B_S) = \frac{1 \pm \epsilon}{|S|}$, where the meaning is clear. Third, the union of i distinct Y_i's has size $n - r + i$. Hence

$$\begin{aligned}
\mathbf{Pr}(B) &= \mathbf{Pr}(B_{Y_1}) + \mathbf{Pr}(B_{Y_2}) + \cdots - \mathbf{Pr}(B_{Y_1} \cap B_{Y_2}) - \cdots + \mathbf{Pr}(B_{Y_1} \cap B_{Y_2} \cap B_{Y_3}) + \cdots \\
&= \mathbf{Pr}(B_{Y_1}) + \mathbf{Pr}(B_{Y_2}) + \cdots - \mathbf{Pr}(B_{Y_1 \cup Y_2}) - \cdots + \mathbf{Pr}(B_{Y_1 \cup Y_2 \cup Y_3}) + \cdots \\
&= \sum_{i=1}^{r} (-1)^{i+1} \binom{r}{i} \frac{1 \pm \epsilon}{n - r + i}
\end{aligned}$$

Hence

$$\begin{aligned}\mathbf{Pr}(B_Y - B) &= \frac{1 \pm \epsilon}{n-r} - \sum_{i=1}^{r}(-1)^{i+1}\binom{r}{i}\frac{1 \pm \epsilon}{n-r+i} \\ &= \sum_{i=0}^{r}(-1)^{i}\binom{r}{i}\frac{1 \pm \epsilon}{n-r+i} \\ &= \sum_{i=0}^{r}(-1)^{i}\binom{r}{i}\frac{1}{n-r+i} \pm \epsilon\sum_{i=0}^{r}\binom{r}{i}\frac{1}{n-r+i}\end{aligned}$$

To evaluate the first term in the expression above, note that it equals $\mathbf{Pr}(B_Y - B)$ when ϵ is 0. That is, the term is the probability that (Z, a) is satisfied for an exactly min-wise independent family. Note that it depends only on n and r, and not on the family under consideration! In particular, we calculate it easily by computing the probability that (Z, a) is satisfied for the family S_n, which is $\frac{1}{(n-r)\binom{n}{r}}$. (Thus we obtain the combinatorial identity

$$\sum_{i=0}^{r}(-1)^{i}\binom{r}{i}\frac{1}{n-r+i} = \frac{1}{(n-r)\binom{n}{r}}.$$

The hint for its algebraic derivation is [21, equation 1.2.6.24].)

The magnitude of the coefficient of ϵ is at most $\frac{2^r}{n-r}$. Hence

$$\frac{1}{(n-r)\binom{n}{r}} + \epsilon\frac{2^r}{n-r} \geq \mathbf{Pr}(B_Y - B) \geq \frac{1}{(n-r)\binom{n}{r}} - \epsilon\frac{2^r}{n-r} \tag{8}$$

Since $\mathbf{Pr}(B_Y - B) \leq \frac{1}{(n-r)\binom{n}{r}} + \epsilon\frac{2^r}{n-r}$, the total probability mass of the permutations that satisfy any given pair (Z, a) is at most $p = \frac{1}{(n-r)\binom{n}{r}} + \epsilon\frac{2^r}{n-r}$. Hence the number of distinct pairs (Z, a) which have some permutation satisfying them must be at least $1/p$. But every permutation satisfies exactly one (Z, a) pair. This means that there must be at least $1/p$ permutations, that is, the size of the family is at least $\frac{(n-r)\binom{n}{r}}{1+\epsilon 2^r\binom{n}{r}}$. □

Corollary 1 *Let $\mathcal{F}$ be exact min-wise independent family, possibly with an associated probability distribution μ. Then $|\mathcal{F}| \geq \lceil \frac{n}{2} \rceil \binom{n}{\lfloor n/2 \rfloor}$.*

Proof: Plug $\epsilon = 0$ and $r = \lfloor \frac{n}{2} \rfloor$ in the result of Theorem 7. □

Remark 5 *Actually, Theorem 7 proves an even stronger corollary: Equation (8) shows that the probability that (Z, a) is satisfied is positive as long as $\epsilon < 1/2^r\binom{n}{r}$. Hence, for any approximate min-wise independent family with such an ϵ, all $\binom{n}{r}(n-r)$ possible pairs (Z, a) are satisfied, and hence there are at least this many permutations. This is maximized for $r = \lfloor \frac{n}{2} \rfloor$, and hence the bound of Corollary 1 also holds for approximate families with an exponentially small ϵ.*

3.5 Lower Bound for Restricted Families

The lower bound of Theorem 7 holds for exactly min-wise independent families. Of course a similar lower bound can also be given for restricted min-wise independent families. For suppose we want the min-wise property to hold for sets of size up to k. Then certainly the property must hold for the set $[k]$, and we may think of all the permutations in our family as acting only on $[k]$. Hence by replacing the value n by k in Theorem 7 we have an appropriate lower bound for restricted min-wise independent families.

Using similar ideas, however, we may achieve better lower bounds on the size of restricted min-wise independent families. Suppose we want the min-wise property to hold for sets of size up to k, and consider any set X such that $|X| \leq k$. For every $X' \subset X, a \in X - X'$, some permutation $\sigma \in \mathcal{F}$ must induce a permutation on X which satisfies (X', a). This means that for some permutation, the only elements of X which occur before a are the elements of X'. Stating this differently, if we split X into disjoint sets X_1,$\{a\}$, and X_2, then there must be some permutation $\sigma \in \mathcal{F}$ such that all the elements of X_1 occur before a and all the elements of X_2 occur after a. Such a permutation is said to *satisfy* the triple (X_1, a, X_2). A triple (X_1, a, X_2) such that $|X_1| + |X_2| + 1 \leq k$, $a \notin X_1$, $a \notin X_2$, and $X_1 \cap X_2 = \emptyset$, is said to be *admissible*. For a restricted min-wise independent family for sets up to size k, every admissible triple must have some permutation satisfying it. This fact is what we use to obtain a lower bound on the number of permutations in the family.

We will focus on admissible triples (X_1, a, X_2) for a fixed $a \in [n]$ and for $|X_1| = |X_2| = \lfloor \frac{k-1}{2} \rfloor$. Let $s = \lfloor \frac{k-1}{2} \rfloor$. We call such triples *symmetric* a-triples. For convenience, assume $a = n-1$. Then $X_1, X_2 \in \binom{[n-1]}{s}$, where this notation denotes that X_1 and X_2 are subsets of size s of the set $[n-1]$.

To obtain our lower bound, we will show that many permutations are needed to satisfy all admissible symmetric a-triples. We do this by associating the set of all symmetric a-triples with the edges of large graph G_a. Similarly, we associate all symmetric a-triples satisfied by a permutation σ with the edges of another, smaller graph $G_{\sigma,a}$. We then show, using the concept of *graph entropy* introduced by Körner [20], that many smaller graphs $G_{\sigma,a}$ are required to cover the edges of the larger graph G_a. This argument will lead to our lower bound.

We now formally define the graphs G_a and $G_{\sigma,a}$. Let $V(G_a) = V(G_{\sigma,a}) = \binom{[n-1]}{s}$; that is, the vertex set contains a vertex corresponding to every s element subset of $[n-1]$. Two vertices are adjacent in G_a if the corresponding sets are disjoint. Every edge in G_a corresponds to a symmetric a-triple. The edge set of $G_{\sigma,a}$ is defined as follows. For $X_1, X_2 \in \binom{[n-1]}{s}$, the edge (X_1, X_2) is present in $G_{\sigma,a}$ if and only if the permutation σ satisfies the triple (X_1, a, X_2). Since every symmetric a-triple must be satisfied by some permutation, for every symmetric a-triple (X_1, a, X_2), the edge (X_1, X_2) must be present in some graph $G_{\sigma,a}$ where $\sigma \in \mathcal{F}$. That is, $\bigcup_{\sigma\in\mathcal{F}} G_{\sigma,a} = G_a$, where here the union is over the edges of the graphs. This fact allows us to obtain a lower bound on the size of $\mathcal{F}$ using graph entropy.

We review briefly the basic facts about graph entropy. We begin with some standard concepts from information theory (see [15].) Note that in what follows we will use X to be a random variable, and not a set as previously, for notational convenience.

Definition 1 (Entropy) *Given a random variable X with a finite range, its* entropy *is given by*

$$H(X) = -\sum_x \mathbf{Pr}[X = x] \log \mathbf{Pr}[X = x]$$

Definition 2 (Mutual Information) *If X and Y are random variables with finite ranges, then their* mutual information *is given by*

$$I(X \wedge Y) = (H(X) - H(X \mid Y)) = H(X) + H(Y) - H((X,Y)).$$

The following definition and results about graph entropy are taken from Körner [20].

Definition 3 (Graph Entropy) *Let $G = (V, E)$ be a graph. Let P be a probability distribution on the vertex set V. Let $\mathcal{A}(G)$ denote the set of all independent sets of G. Let $\mathcal{P}(G)$, the set of admissible distributions, be the set of all distributions Q_{XY} on $V \times \mathcal{A}(G)$ satisfying*

1. *$Q_{XY}(v, A) = 0$ if $v \notin A$, and*
2. *$\sum_A Q_{XY}(v, A) = P(v)$ for all vertices $v \in V$.*

The graph entropy $H(G,P)$ is defined by

$$H(G,P) = \min\{I(X \wedge Y) | Q_{XY} \in \mathcal{P}(G)\}$$

To clarify, in the definition above, X is a random variable representing a vertex of G, and Y is a random variable representing an independent set of G.

Lemma 1 (Sub-additivity of graph entropy) *If G and F are graphs with $V(G) = V(F)$, and P is a distribution on $V(G)$, then $H(F \cup G, P) \leq H(F,P) + H(G,P)$.*

In our discussion, P will always be assumed to be the uniform distribution and will be omitted from our notation for graph entropy. It is easy to see that under this condition, the entropy of the complete graph on n vertices is $\log n$. The entropy of the empty graph is 0. Lemma 1 is central to our lower bound proof. Recall that $\bigcup_{\sigma \in \mathcal{F}} G_{\sigma,a} = G_a$. Thus $H(G_a) \leq \sum_{\sigma \in \mathcal{F}} H(G_{\sigma,a})$. We will show that the entropy of the graphs $G_{\sigma,a}$ is small compared to that G_a. This will give us a lower bound on the size of $\mathcal{F}$.

Lemma 2 (Additivity of Graph Entropy) *Let $\{G_i\}_{i \in I}$ be the set of connected components of a graph G. Then*

$$H(G) = \sum_{i \in I} \frac{|V(G_i)|}{|V(G)|} H(G_i).$$

We state a simple result about the entropy of a complete bipartite graph that we will need later.

Lemma 3 *Let G be a complete bipartite graph on V_1 and V_2, $|V_1| = n_1$ and $|V_2| = n_2$. Then*

$$H(G) \leq p_1 \log \frac{1}{p_1} + p_2 \log \frac{1}{p_2},$$

where $p_1 = \frac{n_1}{n_1+n_2}$ and $p_2 = \frac{n_2}{n_1+n_2}$.

Proof: Let X be a random variable which is uniformly distributed over $V(G) = V_1 \cup V_2$. Let Y be a random variable such that $Y = V_1$ when $X = v$ for $v \in V_1$ and $Y = V_2$ when $X = v$ for $v \in V_2$. With probability p_1, $Y = V_1$ and with probability p_2, $Y = V_2$. Then $H(X) = H((X,Y)) = \log(n_1 + n_2)$. Hence,

$$H(G) \leq H(X) + H(Y) - H((X,Y)) = H(Y) = p_1 \log \frac{1}{p_1} + p_2 \log \frac{1}{p_2}.$$

□

We now compute bounds on the entropies of the graphs G_a and $G_{\sigma,a}$ defined previously.

Lemma 4

$$H(G_a) \geq \log \frac{n-1}{s}$$

Proof: $H(G_a) = H(X) - H(X|Y)$, where X and Y minimize $I(X \wedge Y)$ as in the definition of graph entropy. Recall that X is a random variable that ranges over $V(G_a)$ and Y is a random variable that ranges over $\mathcal{A}(G_a)$, the set of independent sets of G_a. Since the distribution of X is uniform on $V(G_a)$, $H(X) = \log |V(G_a)| = \log \binom{n-1}{s}$. Let $a_{\max}$ be the maximum size of an independent set in G_a. By the Erdős-Ko-Rado theorem (see, for example, [6, Chapter 7]), the maximum size is achieved by the set of vertices corresponding to s element subsets of $[n-1]$ all of which contain some fixed element. Thus $a_{\max} = \binom{n-2}{s-1}$. Now,

$$H(X|Y) = \sum_{A \in \mathcal{A}(G)} H(X|Y = A)\,\mathbf{Pr}(Y = A).$$

For a particular value of Y, say $A \in \mathcal{A}(G_a)$, X is constrained to range over vertices $v \in A$. Thus $H(X|Y = A) \leq \log |A| \leq \log a_{\max}$. Therefore, $H(X|Y) \leq \log a_{\max} = \log \binom{n-2}{s-1}$. This yields

$$H(G_a) \geq \log \binom{n-1}{s} - \log \binom{n-2}{s-1} = \log \frac{n-1}{s}.$$

□

Lemma 5

$$H(G_{\sigma,a}) \leq \frac{1}{2^{s-1}}$$

Proof: Recall that the graph $G_{a,\sigma}$ has an edge (X_1, X_2) for every symmetric a-triple (X_1, a, X_2) satisfied by the permutation σ. Let S_1 be the set of elements that occurs before a in σ and let S_2 be the set of elements that occurs after a in σ. Let $|S_1| = n_1$ and $|S_2| = n_2$, $n_1 + n_2 = n - 1$. Then $G_{\sigma,a}$ has an edge between every set in $\binom{S_1}{s}$ and every set in $\binom{S_2}{s}$. Thus $G_{\sigma,a}$ has a single connected component B of size $\binom{n_1}{s}+\binom{n_2}{s}$. Further, B is a complete bipartite graph and the sizes of its two independent sets are $\binom{n_1}{s}$ and $\binom{n_2}{s}$. By Lemma 3, we have

$$H(B)=\frac{1}{\binom{n_1}{s}+\binom{n_2}{s}}\left[\binom{n_1}{s}\log\left(\frac{\binom{n_1}{s}+\binom{n_2}{s}}{\binom{n_1}{s}}\right)+\binom{n_2}{s}\log\left(\frac{\binom{n_1}{s}+\binom{n_2}{s}}{\binom{n_2}{s}}\right)\right]$$

By Lemma 2, we get

$$\begin{aligned}H(G_{\sigma,a}) &= \frac{|V(B)|}{|V(G_{\sigma,a})|}H(B)\\ &= \frac{1}{\binom{n-1}{s}}\left[\binom{n_1}{s}\log\left(\frac{\binom{n_1}{s}+\binom{n_2}{s}}{\binom{n_1}{s}}\right)+\binom{n_2}{s}\log\left(\frac{\binom{n_1}{s}+\binom{n_2}{s}}{\binom{n_2}{s}}\right)\right]\end{aligned}$$

We provide an upper bound for the expression above. The expression above is equal to $x\log(1+y/x)+y\log(1+x/y)$ for $x=\binom{n_1}{s}/\binom{n-1}{s}$ and $y=\binom{n_2}{s}/\binom{n-1}{s}$. As the function $f(x,y)=x\log(1+y/x)+y\log(1+x/y)$ is increasing in both x and y, we upper bound $H(G_{\sigma,a})$ by $f(x',y')$ where $x'=\left(\frac{n_1}{n-1}\right)^s$ and $x'=\left(\frac{n_2}{n-1}\right)^s$. Now substituting $n_1=\frac{(1+\alpha)}{2}(n-1)$ and $n_2=\frac{(1-\alpha)}{2}(n-1)$ yields

$$H(G_{\sigma,a})\leq\left(\frac{1+\alpha}{2}\right)^s\log\left(1+\left(\frac{1-\alpha}{1+\alpha}\right)^s\right)+\left(\frac{1-\alpha}{2}\right)^s\log\left(1+\left(\frac{1+\alpha}{1-\alpha}\right)^s\right)\quad(9)$$

The right hand side is maximized for $\alpha=0$, i.e. for $n_1=n_2=\frac{n-1}{2}$. Hence the maximum possible value of $H(G_{\sigma,a})$ is $1/2^{s-1}$. □

Since $G_a=\bigcup_{\sigma\in\mathcal{F}}G_{\sigma,a}$, by Lemma 1 $H(G_a)\leq\sum_{\sigma\in\mathcal{F}}H(G_{\sigma,a})$. Hence $|\mathcal{F}|\geq\frac{H(G_a)}{\max_\sigma H(G_{\sigma,a})}\geq 2^{s-1}\log\left(\frac{n-1}{s}\right)$.

The above argument used symmetric a-triples for a fixed value of a. We can give a more careful argument that looks at symmetric a-triples for all values of a. We define auxiliary graphs G_a and $G_{\sigma,a}$ as before. This time, we consider all values of $a\in[n]$. Observe that for each a, $G_a=\bigcup_{\sigma\in\mathcal{F}}G_{\sigma,a}$ and hence $H(G_a)\leq\sum_{\sigma\in\mathcal{F}}H(G_{\sigma,a})$. Summing over all a, we get

$$\begin{aligned}\sum_{a\in[n]}H(G_a) &\leq \sum_{a\in[n]}\sum_{\sigma\in\mathcal{F}}H(G_{\sigma,a})\\ &= \sum_{\sigma\in\mathcal{F}}\sum_{a\in[n]}H(G_{\sigma,a}).\end{aligned}$$

All the graphs G_a are isomorphic and Lemma 4 gives a bound on their entropy. However, for a particular permutation σ, the graphs $G_{\sigma,a}$ are not isomorphic. The proof of Lemma 5 shows that $H(G_{\sigma,a})$ depends on the position of a in permutation σ. It is maximized when a is the middle element of σ and decreases as the distance of a from the middle element increases. The previously computed lower bound used the fact that the maximum entropy of the graphs $G_{\sigma,a}$ is $\frac{1}{2^{s-1}}$. From the bound on $H(G_{\sigma,a})$ in (9) of Lemma 5, we can show that for any fixed permutation σ, the average entropy of the graphs $G_{\sigma,a}$ is $O\left(\frac{1}{s2^s}\right)$. (Here the average is computed over all elements $a \in [n]$.) This yields a lower bound of $\Omega(s2^s \log(\frac{n}{s}))$ on the size of $\mathcal{F}$. Recall that $s = \lfloor\frac{k-1}{2}\rfloor$. Thus we obtain the following theorem.

Theorem 8 *Let $\mathcal{F}$ be any restricted min-wise independent family. Then, $|\mathcal{F}| \geq \Omega(k2^{\frac{k}{2}} \log(\frac{n}{k}))$.*

Also, this entire argument goes through for any approximate, restricted min-wise independent family for sufficiently small ϵ. In fact, we need $\epsilon < \frac{1}{2^s\binom{k}{s}}$ (see Remark 5), and hence $\epsilon < \frac{1}{2^{3k/2}}$ suffices. Thus we have a lower bound of $\Omega(k2^{\frac{k}{2}} \log(\frac{n}{k}))$ for any approximate, restricted min-wise independent family for $\epsilon < \frac{1}{2^{3k/2}}$, i.e. for $k < \frac{2}{3}\log(\frac{1}{\epsilon})$. In general, for given k and ϵ, we take the lower bound for the maximum set size k' such that $k' < \frac{2}{3}\log(\frac{1}{\epsilon})$ and $k' \leq k$. This gives the following lower bound.

Theorem 9 *Let $\mathcal{F}$ be any approximate,restricted min-wise independent family. Then the size of $\mathcal{F}$ is at least*

$$\Omega\left(\min\left(\frac{\log(\frac{1}{\epsilon})(\log n - \log\log(\frac{1}{\epsilon}))}{\epsilon^{\frac{1}{3}}}, k2^{\frac{k}{2}}\log\left(\frac{n}{k}\right)\right)\right).$$

4 Linear and Pairwise Independent Families

We now focus on the behavior of permutations most likely to be used in practice, linear transformations. In particular, we focus on the situation where the universe of elements is $[p]$ for some prime p, and the family of permutations is given by all permutations of the form $\pi(x) = ax + b \bmod p$ (with $a \neq 0$). Linear transformations are easy to represent and efficiently calculable, making them suitable for real applications. Our results suggest that although this family of permutations is not min-wise independent, its performance should be sufficient in many practical situations.

4.1 General Upper and Lower Bounds

As the results for linear permutations require significant calculations, we do not provide proofs for all the results here. We begin with a simple lower bound that holds not just for linear transformations but for any pairwise independent family of permutations; many of our results have this form.

Theorem 10 *For any $X \subseteq [n]$ with $|X| = k$ and for any $x \in X$,*

$$\mathbf{Pr}(\min\{\pi(X)\} = \pi(x)) > \frac{1}{2(k-1)}$$

if π is chosen from a pairwise independent family of permutations.

Proof: Consider a set $X = \{x_0, \ldots x_{k-1}\}$. We will show that $\pi(x_0)$ is the smallest element of $\pi(X)$ as often as required by the theorem. Suppose that $\pi(x_0) = z$. If π is chosen from a pairwise independent family, then $\mathbf{Pr}(\pi(x_i) < z | \pi(x_0) = z) = z/n$. Since the probability that π maps x_i to something smaller than $\pi(x_0)$ is z/n, the probability that π maps any element of X to something smaller than $\pi(x_0)$ is at most $(k-1)z/n$, and hence $\pi(x_0)$ is the minimum of $\pi(X)$ with probability at least $1 - (k-1)z/n$. This is non-negative for $0 \leq z \leq \lfloor \frac{n}{k-1} \rfloor$. Hence

$$\begin{aligned}\mathbf{Pr}(\min\{\pi(X)\} = \pi(x_0)) &\geq \frac{1}{n} \sum_{z=0}^{\lfloor n/(k-1) \rfloor} \left(1 - \frac{(k-1)z}{n}\right) \\ &> \frac{1}{2(k-1)}\end{aligned}$$

□

We have an upper bound on $\mathbf{Pr}(\min\{\pi(X)\} = \pi(x))$ for all pairwise independent families of permutations that is $O(1/\sqrt{k})$, based on a linear programming formulation of the problem. Subsequent to our original proof, Piotr Indyk suggested a simpler proof for this bound [18], so we do not present it here.

4.2 Linear Families, Upper and Lower Bounds

We derive further bounds by considering specifically linear transformations. For instance, we show that the family of linear transformations is not even approximately min-wise independent for any constant ϵ.

Theorem 11 *Consider the set $X_k = \{0, 1, 2 \ldots k\}$, as a subset of $[p]$. As $k, p \to \infty$, with $p \gg k$,*

$$\mathbf{Pr}(\min\{\pi(X)\} = \pi(0)) \sim \frac{3}{\pi^2} \frac{\ln k}{k}$$

when π is a randomly chosen linear transformation of the form $\pi(x) = ax + b \bmod p$ (with $a \neq 0$).

Proof: The proof will use some basic facts about *Farey series*. We first remind the reader of the definition and some basic facts regarding Farey series; more information can be found in most standard number theory texts.

Definition 4 *The* Farey series *of order k consists of all irreducible fractions less than 1 with denominator at most k, in increasing order.*

If $\frac{n_1}{d_1}$ and $\frac{n_2}{d_2}$ are two consecutive fractions in the order k Farey series then

1. $n_2 d_1 - n_1 d_2 = 1$.
2. $(d_1, d_2) = 1$.
3. The first fraction inserted between $\frac{n_1}{d_1}$ and $\frac{n_2}{d_2}$ in a higher order Farey series is $\frac{n_1+n_2}{d_1+d_2}$.

To compute the fraction of time that $\pi(0)$ is the minimum element of $\{\pi(X_k)\}$, let us first consider all transformations π with multiplier a. Let $z_a = \min_{i=1,\ldots,k}\{-a \cdot i \bmod p\}$. Then $\pi(0)$ is minimal only for those $\pi = ax + b \bmod p$ where $b < z_a$ (note that z_a is positive!), since for the other values of b the image of the minimal element will lie behind $\pi(0) = b$.

Hence, to find the fraction of the time that 0 is the minimum element of $\{\pi(X_k)\}$, it suffices to find the expected value of $\frac{1}{p} \min_{i=1,\ldots,k}\{-a \cdot i \bmod p\}$, which conveniently is also the expected value of $\frac{1}{p} \min_{i=1,\ldots,k}\{a \cdot i \bmod p\}$. We concentrate on the latter expression.

Consider what happens to the numbers $\{a \cdot i \bmod p | i = 1 \ldots k\}$ as we increase the value of the multiplier a from 1 to $p - 1$. It is useful to think of the numbers $0, \ldots, p-1$ as arranged clockwise around a circle. Consider k tokens, corresponding to the numbers $1, \ldots, k$ from the set X_k. For each i, we view $a \cdot i \bmod p$ as the position of the ith token at time a. Token i starts in position i. As we increase the value of the multiplier a from 1 to $p - 1$ all tokens move around the circle in clockwise direction but at different speeds: token i moves i steps for every time tick.

If p is sufficiently larger than k, we can think of this motion as being continuous. That is, we scale the circle so that its circumference is 1. Let $f = \frac{a}{p}$. Then the distance of token i from the origin along the circle when the multiplier is a is the fractional part of fi. Henceforth we think of this motion of the tokens as being continuous, with the "time" f increasing uniformly from 0 to 1. We need to compute the average distance of the token closest to the origin as f increases uniformly from 0 to 1, where distance here is measured as clockwise distance along the circumference. This average distance is (asymptotically) $\frac{1}{p}\min_{i=1,\ldots,k}\{a \cdot i \bmod p\}$, the term we wish to compute. (Asymptotically this approximation yields the correct answer, as the approximation affects only lower order terms.)

The token closest to the origin changes whenever a token reaches the origin. This happens whenever the value of f is $\frac{n}{d}$ for integers n and d with $1 \leq n < d \leq k$, as at that point the token with speed d reaches the origin. Thus the times where the token closest to the origin changes are precisely the proper (less than 1) fractions of denominator at most k, that is, the terms of the Farey sequence of order k. Let $\frac{n_1}{d_1}$ and $\frac{n_2}{d_2}$ be two consecutive fractions in the Farey sequence of order k. For $\frac{n_1}{d_1} \leq f \leq \frac{n_2}{d_2}$, the token with speed d_1 is closest to the origin. This time interval has length $\frac{n_2}{d_2} - \frac{n_1}{d_1} = \frac{1}{d_1 d_2}$. During this time interval, the token starts at the origin and moves with a speed of d_1. Thus the average distance of this token from the origin during this interval is $\frac{1}{2} \cdot d_1 \cdot \frac{1}{d_1 d_2} = \frac{1}{2d_2}$.

To obtain the average distance over the entire interval, it suffices to take the appropriate weighted sum over all pairs of consecutive Farey fractions. By the above, the contribution from each interval $[\frac{n_1}{d_1}, \frac{n_2}{d_2}]$ is $\frac{1}{d_1 d_2} \cdot \frac{1}{2d_2} = \frac{1}{2d_1 d_2^2}$.

To find a simple form for the resulting sum, we build up, starting the appropriate sum for $X_1 = \{0, 1\}$ and building up to the set X_k. Alternatively, we may think of how the sum changes as we build up from the order $j-1$ Farey series to the order j Farey series and use this to derive the appropriate sum for the order k Farey series. The order j Farey series is derived from the order $j-1$ Farey series by adding all fractions of the form $\frac{a}{j}$ with $(a, j) = 1$ in their proper position. (Note we use the standard shorthand (a, j) for $\gcd(a, j)$.) Correspondingly, this changes the contribution to the summation in all intervals where a new fraction is inserted. Suppose a fraction is inserted between $\frac{n_1}{d_1}$ and $\frac{n_2}{d_2}$. Then the inserted fraction must be $\frac{n_1+n_2}{d_1+d_2}$, where $d_1 + d_2 = k$. Before the insertion, the contribution of this interval was $\frac{1}{2d_1 d_2^2}$. After the insertion, the contribution becomes $\frac{1}{2d_1(d_1+d_2)^2} + \frac{1}{2(d_1+d_2)d_2^2}$. Thus the change is

$$\frac{1}{2d_1(d_1+d_2)^2} + \frac{1}{2(d_1+d_2)d_2^2} - \frac{1}{2d_1 d_2^2}$$

$$= \frac{d_2^2 + d_1(d_1 + d_2) - (d_1 + d_2)^2}{2d_1(d_1 + d_2)^2 d_2^2}$$
$$= -\frac{1}{2(d_1 + d_2)^2 d_2}$$

Note that $d_1 + d_2 = j$. Further $(j, d_2) = 1$. In fact, for every a such that $(a, j) = 1$, there exists two consecutive Farey fractions $\frac{n_1}{d_1}$ and $\frac{n_2}{d_2}$ such that $d_1 + d_2 = j$ and $d_2 = a$. Thus the change in the summation caused by building up from order $j - 1$ to order j Farey sequences is $-\frac{1}{2j^2} \sum_{(a,j)=1, 1 \leq a \leq j} \frac{1}{a}$. For the order 1 Farey sequence, the value of the appropriate summation is obviously $\frac{1}{2}$. Thus the value for the order k Farey sequence is

$$\frac{1}{2}\left(1 - \sum_{j=2}^{k} \frac{1}{j^2} \sum_{(a,j)=1, 1 \leq a \leq j} \frac{1}{a}\right) \tag{10}$$

From here one must simply evaluate the value of this expression asymptotically to obtain the theorem. This evaluation, unfortunately, requires some work, which we now detail.

First we note that
$$\sum_{j=2}^{\infty} \frac{1}{j^2} \sum_{(a,j)=1, 1 \leq a \leq j} \frac{1}{a} = 1.$$
This follows from the fact that the value for the order k Farey sequence given in (10) must go to 0 as k goes to infinity, since the probability any random point will be the closest to the origin converges to 0. Hence (10) is equivalent to

$$\frac{1}{2} \sum_{j=k+1}^{\infty} \frac{1}{j^2} \sum_{(a,j)=1, 1 \leq a \leq j} \frac{1}{a} = \frac{1}{2} \sum_{j=k+1}^{\infty} \frac{1}{j^3} \sum_{(a,j)=1, 1 \leq a \leq j} \frac{j}{a}.$$

We now employ a common transformation known as Möbius inversion (see, for example, the standard number theory text by Hardy and Wright [17, 16.6.3]). The Möbius inversion yields

$$\frac{1}{2} \sum_{j=k+1}^{\infty} \frac{1}{j^3} \sum_{(a,j)=1, 1 \leq a \leq j} \frac{j}{a} = \frac{1}{2} \sum_{j=k+1}^{\infty} \frac{1}{j^3} \sum_{r|j} \mu\left(\frac{j}{r}\right) \sum_{1 \leq a \leq r} \frac{r}{a}$$

From here we may proceed with straightforward algebraic manipulation. In what follows, we use approximations ($\approx$) in place of equality in expressions where we disregard lower order terms, and we use $H(j) = 1 + 1/2 + \ldots + 1/j$:

$$\begin{aligned}
\frac{1}{2}\sum_{j=k+1}^{\infty}\frac{1}{j^3}\sum_{r|j}\mu\left(\frac{j}{r}\right)rH(r) &= \frac{1}{2}\sum_{j=k+1}^{\infty}\frac{1}{j^3}\sum_{d|j}\mu(d)\frac{j}{d}H\left(\frac{j}{d}\right) \\
&= \frac{1}{2}\sum_{d=1}^{\infty}\mu(d)\sum_{j\geq k+1,d|j}\frac{1}{j^3}\frac{j}{d}H\left(\frac{j}{d}\right) \\
&= \frac{1}{2}\sum_{d=1}^{\infty}\frac{\mu(d)}{d^3}\sum_{i=\lceil\frac{k+1}{d}\rceil}^{\infty}\frac{H(i)}{i^2} \\
&\approx \frac{1}{2}\sum_{d=1}^{\infty}\frac{\mu(d)}{d^3}\int_{k/d}^{\infty}\frac{\ln y}{y^2}dy \\
&\approx \frac{1}{2}\sum_{d=1}^{\infty}\frac{\mu(d)}{d^3}\frac{\ln(k/d)}{(k/d)} \\
&\approx \frac{\ln k}{2k}\sum_{d=1}^{\infty}\frac{\mu(d)}{d^2} \\
&= \frac{\ln k}{2k}\sum_{d=1}^{\infty}\prod_{q\ \text{prime}}\left(1-\frac{1}{q^2}\right) \\
&= \frac{\ln k}{2k}\sum_{d=1}^{\infty}\frac{1}{\sum_{m=1}^{\infty}\frac{1}{m^2}} \\
&= \frac{\ln k}{2k}\frac{6}{\pi^2} = \frac{3}{\pi^2}\frac{\ln k}{k}
\end{aligned}$$

□

Theorem 11 shows that it is possible to find sets for which some element is minimal for $\Omega(\frac{\ln k}{k})$ of the time when random linear transformations are used. Similarly, under linear transformations there is a set X'_k with $k+1$ elements such that $\pi(0)$ is the minimum element of $\pi(X'_k)$ with probability approximately $\frac{12\ln 2}{\pi^2 k} \approx \frac{0.843}{k}$. This result provides an example of how much less often than $\frac{1}{k+1}$ of the time an element can be minimal when random linear transformations are used.

Theorem 12 *Consider the set $X'_k = \{-k/2, \ldots, 0, \ldots k/2\}$, where k is even, as a subset of $[p]$. As $k, p \to \infty$, with $p \gg k$,*

$$\mathbf{Pr}(\min\{\pi(X)\} = \pi(0)) \sim \frac{12\ln 2}{\pi^2 k}$$

when π is a randomly chosen linear transformation of the form $\pi(x) = ax+b \bmod p$ (with $a \neq 0$).

Proof: As before, we think of the numbers as points moving around the circle at different speeds. Here, we have points moving clockwise with speed i for $1 \leq i \leq k/2$, as well as points moving counterclockwise with speed i, for $1 \leq i \leq k/2$. We want to determine the average distance of the point closest to the origin in the clockwise direction; this average distance corresponds to the fraction of the time that $\pi(0)$ is the minimal element of $\pi(X'_k)$.

For a given multiplier a, let $f = \frac{a}{p}$. Let $\frac{n_1}{d_1}$ and $\frac{n_2}{d_2}$ be two consecutive Farey fractions of order $k/2$. During the interval $\frac{n_1}{d_1} \leq f \leq \frac{n_2}{d_2}$, the point moving clockwise with speed d_1 is closest to the origin during the beginning of the interval. It remains so until the time it meets the point moving counterclockwise with speed d_2; this point then remains closest to the origin at the end of the interval. The two points meet at a distance $\frac{1}{d_1+d_2}$ from the origin. The average value distance of the point closest to the origin during this interval is therefore $\frac{1}{d_1 d_2}$ is $\frac{1}{2(d_1+d_2)}$. Hence the contribution of this interval to the overall average value of the minimum is $\frac{1}{2d_1 d_2(d_1+d_2)}$.

As in Theorem 11, to find a simple form for the resulting average distance, we build up by considering the change when we move from the order $j-1$ Farey sequence to the order j Farey sequence. When the fraction $\frac{n_1+n_2}{d_1+d_2}$ is inserted between two consecutive Farey fractions $\frac{n_1}{d_1}$ and $\frac{n_2}{d_2}$, the change in the contribution of the interval $[\frac{n_1}{d_1}, \frac{n_2}{d_2}]$ is

$$\begin{aligned}
&\frac{1}{2}\left[\frac{1}{d_1(d_1+d_2)(2d_1+d_2)} + \frac{1}{d_2(d_1+d_2)(d_1+2d_2)} - \frac{1}{d_1 d_2(d_1+d_2)}\right] \\
&= \frac{d_2(d_1+2d_2) + d_1(2d_1+d_2) - (2d_1+d_2)(d_1+2d_2)}{2d_1 d_2(d_1+d_2)(2d_1+d_2)(d_1+2d_2)} \\
&= -\frac{3}{2(d_1+d_2)(d_1+2d_2)(2d_1+d_2)} \\
&= -\frac{3}{2(d_1+d_2)((d_1+d_2)+d_2)(2(d_1+d_2)-d_2)}
\end{aligned}$$

Note that $d_1 + d_2 = j$ and $(d_2, j) = 1$. Hence the change in the average value in moving from $j-1$ to j is

$$-\frac{3}{2j} \sum_{(a,j)=1, 1 \leq a \leq j} \frac{1}{(j+a)(2j-a)}$$

The value for the order 1 Farey sequence is $\frac{1}{4}$.

Hence the average distance determined by the order $k/2$ Farey sequence is

$$\frac{3}{2}\left(\frac{1}{6}-\sum_{j=2}^{k/2}\frac{1}{j}\sum_{(a,j)=1,1\leq a\leq j}\frac{1}{(j+a)(2j-a)}\right).$$

Using algebraic manipulation similar to that of Theorem 11, one can compute that this summation asympotitically becomes $\frac{12\ln 2}{\pi^2 k}$. □

Despite the seemingly bad worst-case behavior of linear transformations, we believe that in practice they are suitable for applications, because they perform well on random sets. For a set $X=\{x_0,\ldots,x_{k-1}\}$ of size k, let $F(X)$ be $\max_i \frac{|\{\pi \mid \min\{\pi(X)\}=\pi(x_i)\}|}{p(p-1)}$. That is, $F(X)$ is the fraction of the permutations for which the most likely element to be the minimum is actually the minimum. (And we have just seen that $F(X)$ can asymptotically reach $\frac{3}{\pi^2}\frac{\ln k}{k}$ in the worst case.) We now prove that the expected value of $F(X)$ when X is chosen uniformly at random from all sets of size k as $k,p\to\infty$ can be bounded by $(1+1/\sqrt{2})/k+O(1/k^2)$. In this sense, linear transformations are approximately min-wise independent with respect to random sets.

Theorem 13 *As $k,p\to\infty$, with $p\gg k^2$, $\mathbf{E}_X[F(X)]$ is bounded above by $(1+1/\sqrt{2})/k+O(1/k^2)$.*

Proof: We define

$$f_i(X)=\frac{|\{\pi \mid \min\{\pi(X)\}=\pi(x_i)\}|}{p(p-1)},$$

and

$$g_i(z,X)=\frac{|\{\pi \mid \min\{\pi(X)\}=\pi(x_i) \text{ and } \pi(x_i)=z\cdot p\}|}{p-1},$$

That is, consider the subset of permutations that map the ith element to zp. Then g_i is the fraction of these permutations for which the the ith element is minimal.

Hereafter we suppose that the universe size p is sufficiently large that we may think of z as varying continuously on the unit circle from 0 to 1, instead of jumping discretely by $1/p$. This simplification allows us to dismiss many lower order terms. Similarly, we will suppose that p is sufficiently large compared to k so that we may suppose that the k values of X are chosen with replacement, and the results will be equivalent asymptotically. Also, in our calculations, we will find it convenient to replace the $p-1$ term by p in the definitions of $f_i(X)$ and $g_i(z,X)$. Since we are interested in asymptotics as $p\to\infty$, this does not change our results.

The value we wish to bound is

$$F(X) = \mathbf{E}_X[\max_{i=0,\ldots,k-1} f_i(X)],$$

where we use $\mathbf{E}_X$ to denote that the expectation is over the random choice of the set X. Note also that we have the following relation:

$$f_i(X) = \int_0^1 g_i(z, X) dz.$$

Let the $f_i(X)$ have mean μ and variance σ^2. (Note the mean and variance are the same for all f_i.) To bound $F(X)$, we make use of a simple bound on the expected value of the maximum of several identically distributed random variables.

Lemma 6 *Let $X_1, X_2, \ldots, X_k$ be identically distributed random variables with mean μ and variance σ^2. Then*

$$\mathbf{E}[\max_{i=1,\ldots,k} X_i] \leq \mu + \sigma\sqrt{k}.$$

Proof: We show equivalently that

$$\left(\mathbf{E}[\max_{i=1,\ldots,k} X_i - \mu]\right)^2 \leq k\sigma^2.$$

$$\begin{aligned}
\left(\mathbf{E}[\max_{i=1,\ldots,k} X_i - \mu]\right)^2 &\leq \mathbf{E}\left[(\max_{i=1,\ldots,k} X_i - \mu)^2\right] \\
&\leq \mathbf{E}\left[\max_{i=1,\ldots,k} (X_i - \mu)^2\right] \\
&\leq \mathbf{E}\left[\sum_{i=1,\ldots,k} (X_i - \mu)^2\right] \\
&= \sum_{i=1,\ldots,k} \mathbf{E}[(X_i - \mu)^2] \\
&= k\sigma^2
\end{aligned}$$

□

Clearly, by symmetry $\mathbf{E}_X[f_i(X)] = 1/k$. Hence, to find an upper bound on F, we just have to bound σ^2, the variance of $f_i(X)$. Specifically, we bound the variance of $f_0(X)$.

We define some helpful notation. Let $\pi_{a,z}$ denote the unique linear permutation such that $ax_0+b = z\cdot p \bmod p$. That is, $\pi_{a,z}$ is the linear permutation with multiplier a that maps x_0 to $z\cdot p$. Let $M_a(z,X)$ be an indicator random variable that is 1 if $\min\{\pi_{a,z}(X)\} = \pi_{a,z}(x_0)$. Thus, $g_0(z,X) = \frac{1}{p}\sum_a M_a(z,X)$. Now the variance of f_0 is just

$$\begin{aligned}
\sigma^2 &= \mathbf{E}_X\left[(f_0(X) - \mathbf{E}_X[f_0(X)])^2\right]\\
&= \mathbf{E}_X\left[\left(\int_0^1 g_0(z,X)dz - \mathbf{E}_X\left[\int_0^1 g_0(z,X)dz\right]\right)^2\right]\\
&= \mathbf{E}_X\left[\left(\int_0^1 \left(g_0(z,X) - E_X\left[g_0(z,X)\right]\right)dz\right)^2\right]\\
&= \frac{1}{p^2}\mathbf{E}_X\Bigg[\Bigg(\int_0^1\Bigg(\sum_a M_a(z,X)\\
&\qquad\qquad - E_X\Bigg[\sum_a M_a(z,X)\Bigg]\Bigg)dz\Bigg)^2\Bigg]\\
&= \frac{1}{p^2}\mathbf{E}_X\Bigg[\Bigg(\int_0^1\sum_a\Bigg(M_a(z,X)\\
&\qquad\qquad - E_X\Big[M_a(z,X)\Big]\Bigg)dz\Bigg)^2\Bigg]
\end{aligned}$$

Let $\mu_a(z) = E_X\left[M_a(z,X)\right]$. From this definition, it is apparent that $\mu_a(z) = (1-z)^{k-1}$, as each of the images of the other randomly chosen $k-1$ elements has probability $1-z$ of being greater than $z\cdot p$.

Hence, continuing from the last line above,

$$\begin{aligned}
\sigma^2 &= \frac{1}{p^2}\mathbf{E}_X\Bigg[\Bigg(\int_0^1\sum_a(M_a(z,X) - E_X\left[M_a(z,X)\right])dz\Bigg)^2\Bigg]\\
&= \frac{1}{p^2}\mathbf{E}_X\Bigg[\int_{z=0}^1\int_{y=0}^1\Bigg(\sum_{a_1,a_2}(M_{a_1}(z,X) - \mu_{a_1}(z))\\
&\qquad\qquad \times (M_{a_2}(y,X) - \mu_{a_2}(y))\Bigg)dy\,dz\Bigg]\\
&= \frac{1}{p^2}\int_{z=0}^1\int_{y=0}^1\Bigg(\sum_{a_1,a_2}\Bigg(\mathbf{E}_X\left[M_{a_1}(z,X)M_{a_2}(y,X)\right]
\end{aligned}$$

$$- \mu_{a_1}(z)\mu_{a_2}(y)\Big)\Big) dy\, dz \tag{11}$$

We now bound the last term. This will in turn bound the variance and yield the theorem. In order to do this, we derive an alternative expression for $\mathbf{E}_X[M_{a_1}(z,X)M_{a_2}(y,X)]$ that can be be appropriately bounded.

Let

$$q_{a_1,a_2}(z,y) = \mathbf{Pr}_{x\in[p]}(\pi_{a_1,z}(x) > z\cdot p \text{ and } \pi_{a_2,y}(x) > y \cdot p).$$

Then

$$\mathbf{E}_X[M_{a_1}(z,X)M_{a_2}(y,X)] = \left(q_{a_1,a_2}(z,y)\right)^{k-1},$$

again since the other $k-1$ terms of X are chosen uniformly at random.

We thus have expressed the value we wish to bound as the sum of the $(k-1)$st powers of q_{a_1,a_2} terms. The next lemma shows that the sum of these q_{a_1,a_2} terms is fixed. As the maximum possible value of the sum of the $(k-1)$st powers is achieved when the terms in the sum take on extremal values, together these results will allow us to bound $\sum_{a_1,a_2}\mathbf{E}_X[M_{a_1}(z,X)M_{a_2}(y,X)]$.

Lemma 7

$$\sum_{a_1,a_2} q_{a_1,a_2}(z,y) = p^2(1-z)(1-y).$$

Proof: Consider the following experiment. We choose three values $a_1, a_2, x \in [p]$ independently and uniformly at random. The experiment succeeds if both $\pi_{a_1,z}(x) > z\cdot p$ and $\pi_{a_2,y}(x) > y\cdot p$. Clearly, the probability of success is $(1-z)(1-y)$. The summation $\sum_{a_1,a_2} p\cdot q_{a_1,a_2}(z,y)$ is simply the number of the p^3 triples (a_1,a_2,x) for which the experiment succeeds. The lemma follows. □

Since the total sum of the terms q_{a_1,a_2} is fixed, the sum $\sum_{a_1,a_2}\mathbf{E}_X[M_{a_1}(z,X)M_{a_2}(y,X)]$ is maximized when the q_{a_1,a_2} terms take on extremal values. Let us assume that $z \geq y$. Then $q_{a_1,a_2}(z,y) \in [1-z-y, 1-z]$. (Of course $q_{a_1,a_2}(z,y) \geq 0$, and hence the above range may not be correct if $z+y > 1$.) A simple calculation then yields the following bound (for $z+y \leq 1$):

$$\sum_{a_1,a_2}\mathbf{E}_X[M_{a_1}(z,X)M_{a_2}(y,X)] \\ \leq p^2\left(z(1-z)^{k-1} + (1-z)(1-z-y)^{k-1}\right).$$

We will use this bound for the range $z \leq 1/2$. For $z > 1/2$, we have $q_{a_1,a_2}(z,y) \leq 1 - z \leq 1/2$. Hence,

$$\sum_{a_1,a_2} \mathbf{E}_X[M_{a_1}(z,X)M_{a_2}(y,X)] \leq p^2(1/2^{k-1}).$$

Substituting this bound in (11), we get:

$$\begin{aligned}
\sigma^2 &= \frac{1}{p^2}\,\mathbf{E}_X\Big[\int_{z=0}^{1}\int_{y=0}^{1}\Big(\sum_{a_1,a_2}\Big(M_{a_1}(z,X)M_{a_2}(y,X) \\
&\qquad\qquad - \mu_{a_1}(z)\mu_{a_2}(y)\Big)\Big)dy\,dz\Big] \\
&= \frac{2}{p^2}\int_{z=0}^{1}\int_{y=0}^{z}\Big(\sum_{a_1,a_2}\mathbf{E}_X\Big[M_{a_1}(z,X)M_{a_2}(y,X) \\
&\qquad\qquad - \mu_{a_1}(z)\mu_{a_2}(y)\Big]\Big)dy\,dz \\
&\leq 2\int_{z=0}^{1/2}\int_{y=0}^{z}\Big(z(1-z)^{k-1} + (1-z)(1-z-y)^{k-1} \\
&\qquad\qquad - (1-z)^{k-1}(1-y)^{k-1}\Big)dy\,dz \\
&\quad + 2\int_{z=1/2}^{1}\int_{y=0}^{z}\frac{1}{2^{k-1}}\,dy\,dz
\end{aligned}$$

To prove Theorem 1, we need merely to compute this integral thus bounding the variance. This calculation is easily performed, yielding

$$\sigma^2 \leq \frac{1}{2k^3} + O(1/k^4).$$

This proves Theorem 13. □

Simulations suggest that in fact the behavior of families of linear transformations on a random set X is much better than this. We conjecture that the expected value of $F(X)$ converges to $1/k$ asymptotically.

Also, we note that Theorem 13 actually generalizes quite straightforwardly to all pairwise independent families. The notation becomes slightly more difficult, as one must take care to index variables and summations appropriately, but the proof follows the same course.

5 Acknowledgments

The authors thank Noam Elkies for enlightening discussions regarding Farey series.

References

[1] N. Alon, M. Dietzfelbinger, P. B. Miltersen, E. Petrank, and G. Tardos. Is linear hashing good? In *Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing*, pages 465–474, El Paso, Texas, 4–6 May 1997.

[2] N. Alon, O. Goldreich, J. Håstad, and R. Peralta. Simple constructions of almost k-wise independent random variables. *Random Structures and Algorithms*, 3(3):289–304, 1992.

[3] N. Alon and J. H. Spencer. *The Probabilistic Method.* John Wiley and Sons, 1992.

[4] T. M. Apostol. *Introduction to Analytic Number Theory.* Springer-Verlag, 1976.

[5] T. Berners-Lee, R. Cailliau, A. Loutonen, H. F. Nielsen, and A. Secret. The world-wide web. *Communications of the ACM*, 37(8):76–82, 1994.

[6] B. Bollobás. *Combinatorics: Set Systems, Hypergraphs, Families of Vectors, and Combinatorial Probability.* Cambridge University Press, 1986.

[7] A. Z. Broder. On the resemblance and containment of documents. In *Proceedings of Compression and Complexity of SEQUENCES 1997.* To appear.

[8] A. Z. Broder. Filtering near-duplicate documents. In *Proceedings of FUN 98*, 1998. To appear.

[9] A. Z. Broder, M. Charikar, A. Frieze, and M. Mitzenmacher. Min-wise independent permutations. In *Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing*, pages 327–336, 1998.

[10] A. Z. Broder, M. Charikar, and M. Mitzenmacher. Derandomization using min-wise independent permutations. Draft manuscript.

[11] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the Web. In *Proceedings of the Sixth International World Wide Web Conference*, pages 391–404, 1997.

[12] A. Z. Broder and A. R. Karlin. Multilevel adaptive hashing. In *Proceedings of the First Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 43–53, San Francisco, California, 22–24 Jan. 1990.

[13] J. L. Carter and M. N. Wegman. Universal classes of hash functions. *Journal of Computer and System Sciences*, 18(2):143–154, Apr. 1979.

[14] E. Cohen. Estimating the size of the transitive closure in linear time. In *Proceedings of the Thrity-Fifth Annual IEEE Symposium on Foundations of Computer Science*, pages 190–200, 1994.

[15] T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley & Sons, Inc., 1991.

[16] M. Dietzfelbinger, A. Karlin, K. Mehlhorn, F. M. auf der Heide, H. Rohnert, and R. E. Tarjan. Dynamic perfect hashing: Upper and lower bounds. *SIAM J. Comput.*, 23(4):738–761, Aug. 1994.

[17] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers.* Oxford University Press, fifth edition, 1979.

[18] P. Indyk. Personal communication.

[19] 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. In *Proceedings of the Twenty-Ninth Annual ACM Symposium on the Theory of Computing*, pages 654–663, El Paso, Texas, May 1997.

[20] J. Körner. Fredman-Komlós bounds and information theory. *SIAM J. Alg. Disc. Meth.*, 7(4):560–570, 1986.

[21] D. E. Knuth. *The Art of Computer Programming, Vol. I: Fundamental Algorithms.* Addison-Wesley, second edition, 1973.

[22] M. Luby and A. Wigderson. Pairwise independence and derandomization. Technical Report TR-95-035, International Computer Science Institute, Berkeley, California, 1995.

[23] R. Seltzer, E. J. Ray, and D. S. Ray. *The Alta Vista Search Revolution : How to Find Anything on the Internet.* McGraw-Hill, 1996.

[24] R. J. Souza, P. Krishnakumar, C. M. Özveren, R. J. Simcoe, B. A. Spinney, R. E. Thomas, and R. J. Walsh. GIGAswitch: A high-performance packet-switching platform. *DIGITAL Technical Journal*, 6(1):9–22, 1994.

[25] D. Zuckerman. Personal communication.