SIAM J. COMPUT.
Vol. 37, No. 1, pp. 83–111

A PROBABILISTIC STUDY ON COMBINATORIAL EXPANDERS AND HASHING*

PHILLIP G. BRADFORD† AND MICHAEL N. KATEHAKIS‡

Abstract. This paper gives a new way of showing that certain constant degree graphs are graph expanders. This is done by giving new proofs of expansion for three permutations of the Gabber–Galil expander. Our results give an expansion factor of $\frac{3}{16}$ for subgraphs of these three-regular graphs with $(p-1)^2$ inputs for p prime. The proofs are not based on eigenvalue methods or higher algebra. The same methods show the expected number of probes for unsuccessful search in double hashing is bounded by $\frac{1}{1-\alpha}$, where α is the load factor. This assumes a double hashing scheme in which two hash functions are randomly and independently chosen from a specified uniform distribution. The result is valid regardless of the distribution of the inputs. This is analogous to Carter and Wegman's result for hashing with chaining. This paper concludes by elaborating on how any sufficiently sized subset of inputs in any distribution expands in the subgraph of the Gabber–Galil graph expander of focus. This is related to any key distribution having expected $\frac{1}{1-\alpha}$ probes for unsuccessful search for double hashing given the initial random, independent, and uniform choice of two universal hash functions.

Key words. expander graphs, double hashing, Gabber–Galil expander, expansion factor, combinatorial expanders, pairwise independence, hash collisions

AMS subject classifications. 05C90, 68P05, 68P20, 60C05

DOI. 10.1137/S009753970444630X

1. Introduction.

Consider a bipartite graph $G = (\mathcal{I} \cup \mathcal{O}, E)$, where $\mathcal{I}$ and $\mathcal{O}$ form the bipartition of the nodes with $n = |\mathcal{I}| = |\mathcal{O}|$, and let r be G's maximum degree. G is an (n, r, c)-*expander* if the following holds:

$$|\mathcal{N}(\widehat{\mathcal{I}})| \geq |\widehat{\mathcal{I}}| \left(1 + c\left(1 - \frac{|\widehat{\mathcal{I}}|}{n}\right)\right),$$

for every subset $\widehat{\mathcal{I}} \subset \mathcal{I}$ that contains up to $n/2$ elements (inputs), where $\mathcal{N}(\widehat{\mathcal{I}}) = \{o \in \mathcal{O} : (i, o) \in E \text{ for some } i \in \widehat{\mathcal{I}}\}$. The constant c is the expansion factor of the graph. Further, the degree r is bounded by a constant. There are many other essentially equivalent definitions for graph expanders. The "hard part" in designing graph expanders is proving they expand. In fact, the decision problem of determining expansion is co-$\mathcal{NP}$-complete [6].

A series of classic papers firmly established that certain graph families expand. First, Margulis [21] showed that expanders exist, without giving bounds on their expansion. However, he did show how to construct them explicitly. Next, Gabber and Galil [12] gave an explicit expander construction with bounds on their expansion.

*Received by the editors October 15, 2004; accepted for publication (in revised form) November 1, 2006; published electronically March 30, 2007. This research was supported by the Rutgers Business School Research Resources Committee.

http://www.siam.org/journals/sicomp/37-1/44630.html

†Department of Computer Science, University of Alabama, Box 870290, Tuscaloosa, AL 35487-0290 (pgb@cs.ua.edu).

‡Department of Management Sciences and Information Systems, Rutgers Business School—Newark and New Brunswick, Rutgers University, 180 University Ave., Newark, NJ 07102 (mnk@andromeda.rutgers.edu).

Finally, Alon [2] showed that a graph is an expanding graph iff its largest and second-largest eigenvalues are well separated. See also, for example, [5, 4, 26, 10, 18, 29] for varying depths of coverage of eigenvalue methods for graph expansion. The eigenvalue methods have been central in much research on graph expanders.

Eigenvalue methods do not give the best possible expanding graph coefficients [33]. For example, probabilistic methods show the existence of expanders that have better expansion than is possible to show by the separation of the largest and second-largest eigenvalues. Pinsker [27] first showed the existence of expanders using probabilistic methods.

There are some other constructions of expanders. According to Alon [3], the (eigenvalue-based) construction of Jimbo and Maruoka [15] "only uses elementary but rather complicated tools from linear algebra." Ajtai [1] also gives an algorithm using linear algebra for constructing three-regular expanding graphs. This algorithm is complex and takes $O(n^3 \log^3 n)$ time to construct an expander. The expansion factor of these expanders is unknown but positive. Lubotzky, Phillips, and Sarnak [19] and independently Margulis [22] gave the best possible expanders using the eigenvalue methods [2, 18, 19, 29]. Kahale [16] gave the best expansion constant to date for Ramanujan and related graphs. Reingold, Vadhan, and Wigderson [28] give very important combinatorial constructions of constant degree expanders based on their new "zig-zag" graph product. By showing how the zig-zag product maintains the eigenvalue bounds (then breaks them), they show how to construct expanders recursively starting from a small expander. Further, Meshulam and Wigderson [25] give group theoretic techniques whose expansion they show depends on universal hash functions. Capalbo et al. [7] give constant degree d lossless expanders. These expand by $(1-\epsilon)d$, for $\epsilon > 0$, which is just about as much as possible.

We demonstrate expansion of $\frac{3}{16} = 0.1875$ for three of the five permutations that comprise the Gabber–Galil expander [12]. These results hold for three-regular subgraphs of the Gabber–Galil graphs of p^2 input vertices, where p is a prime. This is done without using Eigenvalue based bounds. The actual Gabber–Galil expansion was shown to be $(2-\sqrt{3})/4$ or about 0.067.

Suppose double hashing is based on randomly, independently, and uniformly choosing two hash functions h_1 and h_2 from a universal set [11]. Then this paper shows the expected number of probes for unsuccessful search in double hashing is bounded by $\frac{1}{1-\alpha}$, where α is the load factor. This holds regardless of the distribution of the inputs. This is analogous to Carter and Wegman's result for hashing with chaining.

1.1. Intuitive overview. Given three permutations of the Gabber–Galil expander graph, this paper shows no matter what subset of inputs (up to half of them) an adversary chooses, then there is at least $\frac{3}{16}$ expansion. This is done in two steps while trading off the local and global structure of the graph. If the adversary allows enough local expansion, then we are done. Therefore, assume the adversary focuses on sufficiently restricting local expansion. In this case, the adversary must choose inputs in certain patterns. Now, in the second step of our main result, it is shown that these patterns cannot block much global expansion.

If the elements are in the appropriate local patterns to minimize local expansion, then the adversary has freedom to choose the number of elements in the patterns as well as where these patterns start. Certain global patterns are *collision sequences* (see Definition 3). Collision sequences reduce the global expansion. Constraining ourselves to local input patterns, the expected length of all of these collision sequences is at

most 2, no matter how the adversary chooses to position the local patterns or how many elements the adversary chooses to put in them.

It is essential to note that showing the expected collision sequence length is at most 2 uses probability theory applied to the adversary's constrained selections of input node patterns. Our argument shows the adversary has some very restricted choices of input nodes in the three fixed permutations of Gabber–Galil's graph; otherwise, the adversary allows lots of local expansion. At all times, the three permutations comprising the Gabber–Galil graph remain fixed. The results are given by using probabilistic methods on these fixed graphs.

Further, using virtually the same methods, start by randomly, uniformly, and independently selecting two universal hash functions h_1 and h_2 to build a double hashing table T. All elements will be put in T by double hashing using h_1 and h_2. In this case, let T have fixed load factor $\alpha : 1 > \alpha > 0$. Then we show the expected number of probes for an unsuccessful search in T, still using these initially chosen hash functions, is $\frac{1}{1-\alpha}$. As in the case of our expander result, we show this using probabilistic techniques on fixed graphs.

1.2. Structure of this paper. Section 2 gives details of the three permutations comprising the Gabber–Galil expander and sets the foundations for showing both expansion as well as our hashing result. Section 2 has five subsections. Subsection 2.1 gives the actual graph construction. Next, subsection 2.2 defines local and global expansion. Subsection 2.3 explains the relation of double hashing to the expander graph representation. Next, subsection 2.4 focuses on the results of Chor and Goldreich [9] showing randomly choosing such functions and computing their values gives pairwise independent and uniformly distributed values. Finally, subsection 2.5 bounds functions that are necessary for our final result.

Section 3 uses our methods to show that randomly independently and uniformly selecting two double hash functions from a strongly universal set gives a double hashing result analogous to the classical result of Carter and Wegman [8] for hashing with chaining.

Section 4 completes the expander result, showing the subgraphs expand by $\frac{3}{16}$ by enunciating the trade-off of local and global expansion. Finally, in section 5 we give our conclusions and tie together the notion of expansion with the notion of double hashing with universal hash functions.

2. Combinatorial expanders. This section gives the construction and starts the analysis of expanders without using eigenvalue bounds. Without loss, always assume that $n = |\mathcal{I}| = |\mathcal{O}|$ and $n = p^2$, where p is a prime. Let $\widehat{\mathcal{I}}$ denote the elements from $\mathcal{I}$ that an adversary selects from $\mathcal{I}$ in trying to foil any expansion. The adversary foils an expansion by selecting inputs in such a way so there are relatively few outputs. This section shows that no matter what set $\widehat{\mathcal{I}}$ the adversary chooses, there is expansion.

2.1. The construction. This subsection constructs three-regular bipartite graphs $G_3 = (V, E)$ with vertices $V = \mathcal{I} \cup \mathcal{O}$ denoting the *inputs* and *outputs*, respectively. This graph is made up of permutations σ_0, σ_2, and σ_3 used in building Gabber and Galil's expander [12]. The permutations comprising the Gabber–Galil expander are very similar to the permutations that make up Margulis' expander.

Only inputs can have edges to outputs. Let $\mathbb{Z}_p^+ = \{0, 1, \ldots, p-1\}$. Partition the inputs $\mathcal{I}$ and the outputs $\mathcal{O}$ into p blocks I_j and O_j, for all $j \in \mathbb{Z}_p^+$, containing p

nodes each. In particular, for any $j : p > j \geq 0$,

$$I_j = \{\ (j,0),\ (j,1),\ \ldots,\ (j,p-1)\ \},$$
$$O_j = \{\ (j,0)',\ (j,1)',\ \ldots,\ (j,p-1)'\ \}.$$

For notational convenience let (j,k) denote the kth element of both lists I_j and O_j for all $j,k \in \mathbb{Z}_p^+$.

As an example, consider $p = 3$ in Figure 1.

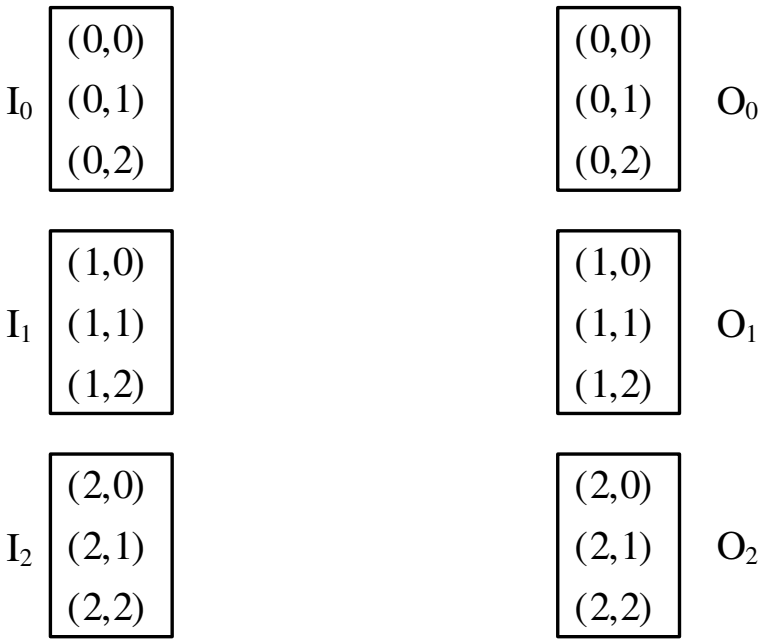

FIG. 1. *The nodes in G_3 where $p = 3$.*

Now take $\mathcal{I}$ as

$$\mathcal{I} = \bigcup_{j=0}^{p-1} I_j.$$

Likewise, for $\mathcal{O}$,

$$\mathcal{O} = \bigcup_{j=0}^{p-1} O_j.$$

For any input node $(j,k) \in I_j$ such that $j \in \mathbb{Z}_p^+$ and $k \in \mathbb{Z}_p^+$, the graph G_3 has the following edges:

1. *Identity*: $\boldsymbol{id}(j,k) \longrightarrow (j,k)$.
2. *Local shift*: $\boldsymbol{loc}(j,k) \longrightarrow (j,\ (j+k+1) \bmod p)$.
3. *Global shift*: $\boldsymbol{g}(j,k) \longrightarrow ((j+k) \bmod p,\ k)$.

These edges are directed from the inputs to the outputs. This does not affect the expansion since it is measured from how the inputs expand to the outputs. Likewise, these directed edges are consonant with the hashing result given in this paper.

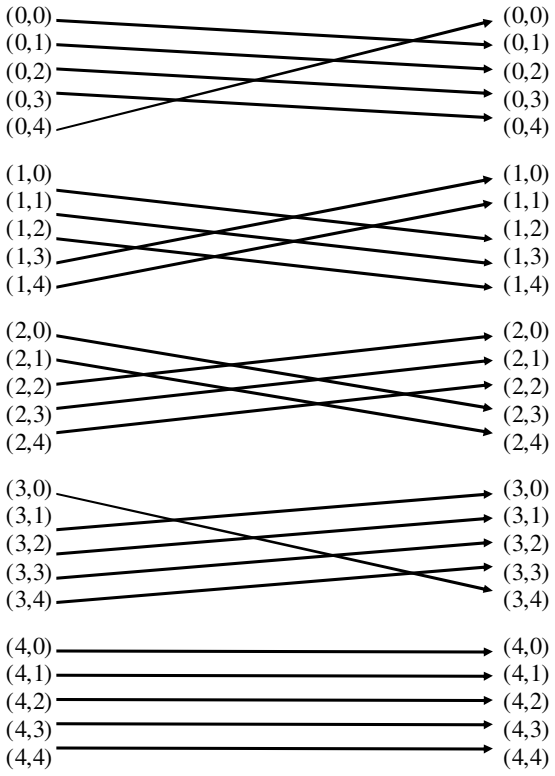

FIG. 2. *Local edges in G_3 where $p = 5$.*

Figure 2 gives local edges for G_3 where $p = 5$, and Figure 3 gives global edges for input blocks I_0 and I_1 in G_3 where $p = 5$. The identity edges are not shown in either of these figures. Also see Gabber and Galil [12] or, for example, Motwani and Raghavan [26]. Note in block I_{p-1} the local shift edges degenerate as $\boldsymbol{loc}(p-1,k) = (p-1,k)$ for all $k \in \mathbb{Z}_p^+$. Likewise, in nodes $(j,0)$ the global shift edges degenerate as $\boldsymbol{g}(j,0) = (j,0)$ for all $j \in \mathbb{Z}_p^+$. Therefore, these nodes $(j,0)$ for $\boldsymbol{g}$ and $(p-1,k)$ for $\boldsymbol{loc}$ do not share all of the necessary properties for expansion. Generally, this paper assumes the adversary does not select these degenerate elements. However, after the main theorems, Theorems 8 and 9, an accounting is made assuming the adversary does select degenerate elements.

These maps are well defined on sets. So $\boldsymbol{id}(S) \cup \boldsymbol{loc}(S) \cup \boldsymbol{g}(S) = \mathcal{N}(S) \subseteq \mathcal{O}$ for any set of inputs $S \subseteq \mathcal{I}$. Further, $\boldsymbol{g}_1(j,k), \boldsymbol{loc}_1(j,k)$, and $\boldsymbol{id}_1(j,k)$ denote the first component of the pair, while $\boldsymbol{g}_2(j,k), \boldsymbol{loc}_2(j,k)$, and $\boldsymbol{id}_2(j,k)$ denote the second element. An instance of this subcase of the Gabber–Galil expander is

$$G_3 = (\mathcal{O} \cup \mathcal{I}, \boldsymbol{id}(\mathcal{I}) \cup \boldsymbol{loc}(\mathcal{I}) \cup \boldsymbol{g}(\mathcal{I})).$$

2.2. The analysis. An adversary, who knows G_3's construction, selects sublists $\widehat{I_j}$ from each block I_j. A sublist may be empty. This paper shows that no matter what elements the adversary selects, the graph G_3 expands. This paper assumes up

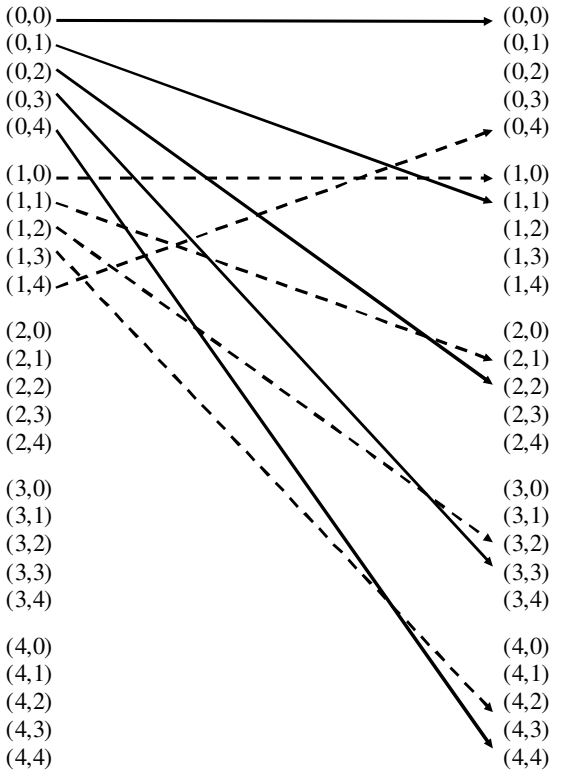

FIG. 3. *Global edges in G_3, from I_0 and I_1, where $p = 5$.*

to half of the inputs to be chosen by the adversary:

$$\sum_{j=0}^{p-1} |\widehat{I}_j| \leq \left\lfloor \frac{n}{2} \right\rfloor .$$

Let

$$\widehat{\mathcal{I}} = \bigcup_{j=0}^{p-1} \widehat{I}_j.$$

DEFINITION 1. *Given block I_j, for $j \in \mathbb{Z}_p^+$, the local $\mathcal{L}$ and global $\mathcal{G}$ expansions of I_j are*

$$\mathcal{L}(\widehat{I}_j) = |\boldsymbol{loc}(\widehat{I}_j) - \boldsymbol{id}(\widehat{I}_j)|,$$
$$\mathcal{G}(\widehat{I}_j) = |\boldsymbol{g}(\widehat{\mathcal{I}}) \cap O_j - \boldsymbol{id}(\widehat{I}_j)|.$$

Definition 1 immediately gives

$$\begin{aligned} \mathcal{L}(\widehat{\mathcal{I}}) &= \sum_{j=0}^{p-1} \mathcal{L}(\widehat{I}_j) \\ &= |\boldsymbol{loc}(\widehat{\mathcal{I}}) - \boldsymbol{id}(\widehat{\mathcal{I}})| \end{aligned}$$

and

$$\begin{aligned}\mathcal{G}(\widehat{\mathcal{I}}) &= \sum_{j=0}^{p-1} \mathcal{G}(\widehat{I}_j) \\ &= |\boldsymbol{g}(\widehat{\mathcal{I}}) - \boldsymbol{id}(\widehat{\mathcal{I}})|.\end{aligned}$$

Local and global expansion can "collide" in that output nodes that give local expansion can also give global expansion. That is, there may be some $\widehat{\mathcal{I}}' \subseteq \widehat{\mathcal{I}}$, where $\boldsymbol{loc}(\widehat{\mathcal{I}}') = \boldsymbol{g}(\widehat{\mathcal{I}}')$. In this case, to compute the total expansion of $\widehat{\mathcal{I}}'$ just divide $\mathcal{L}(\widehat{\mathcal{I}}') + \mathcal{G}(\widehat{\mathcal{I}}')$ by 2. Likewise, if local expansion and global expansion share output nodes, then just consider the case that offers more expansion (if they do not offer the same expansion).

For ease of exposition, when possible we generally refer to the elements of the inputs $\mathcal{I}$ from here on. Each input is directly associated with the element that it maps to by the identity mapping.

DEFINITION 2. *In a block $\widehat{I}_j$, for some $j \in \mathbb{Z}_p^+$, the element $(j, k_1) \in \widehat{I}_j$ is* $\boldsymbol{loc}$-contiguous *iff $\boldsymbol{loc}(j, k_1) = \boldsymbol{id}(j, k_2)$ for $k_2 = (j + k_1 + 1) \bmod p$ and $(j, k_2) \in \widehat{I}_j$. A* $\boldsymbol{loc}$-contiguous set *is a list $(j, k_1), (j, k_2), \ldots, (j, k_t)$ all in $\widehat{I}_j$ and $\boldsymbol{loc}(j, k_s) = \boldsymbol{id}(j, k_{s+1})$ for all $s : t > s \geq 1$.*

If $\mathcal{L}(\widehat{I}_j) \leq 1$, for some $j \in \mathbb{Z}_p^+$, then the elements in $\widehat{I}_j$ are $\boldsymbol{loc}$-contiguous.

LEMMA 1. *If there exists a fixed $d : 1 \geq d > 0$, where $d|\widehat{I}_j| \geq |\boldsymbol{id}(\widehat{I}_j) \cap \boldsymbol{loc}(\widehat{I}_j)|$, for all blocks I_j such that $j \in \mathbb{K} \subseteq \mathbb{Z}_p^+$, where $\mathbb{K} \neq \emptyset$ and $\widehat{I}_j \neq \emptyset$, then $\mathcal{L}(\widehat{\mathcal{I}}) \geq (1 - d) \sum_{j \in \mathbb{K}} |\widehat{I}_j|$.*

Proof. First, since $|\boldsymbol{id}(\widehat{I}_j) \cap \boldsymbol{loc}(\widehat{I}_j)| \leq d|\widehat{I}_j|$ so $|\boldsymbol{loc}(\widehat{I}_j) - \boldsymbol{id}(\widehat{I}_j)| \geq (1 - d)|\widehat{I}_j|$, therefore it must be that $\mathcal{L}(\widehat{I}_j) \geq (1-d)|\widehat{I}_j|$. Since $\boldsymbol{loc}(\widehat{I}_j) \subseteq I_j$, this proof generalizes for the index set $\mathbb{K}$. □

DEFINITION 3. *A* collision sequence *of* length *t is the maximal sequence of elements $(j_1, k), \ldots, (j_t, k)$, where $t \geq 1$, such that $(j_i, k) \in \widehat{I_{j_i}}$, for all $i \in \{1, \ldots, t\}$ and $\boldsymbol{g}(j_0, k) \notin \widehat{I}_{j_0}$ and $(j_{t+1}, k) \notin \widehat{I}_{j_{t+1}}$, where*

$$\begin{aligned}\boldsymbol{g}(j_0, k) &\longrightarrow (j_1, k), \\ \boldsymbol{g}(j_1, k) &\longrightarrow (j_2, k), \\ &\vdots \\ \boldsymbol{g}(j_t, k) &\longrightarrow (j_{t+1}, k).\end{aligned}$$

So $\mathbf{Length}((j_1, k)) = t$.

For example, if $(j_1, k) \in \widehat{I}_{j_1}$, but $\boldsymbol{g}(j_1, k) \notin \widehat{I}_t$, where $t = \boldsymbol{g}_1(j_1, k)$, then (j_1, k) is a length 1 collision sequence starting in input block I_{j_1}. Therefore, a collision sequence of length 1 starts and ends in the same block. Length 1 collision sequences do not diminish expansion but rather increase it.

DEFINITION 4. *Suppose the elements $(j_s, k) \in \widehat{I}_{j_s}$ for all $s : t \geq s \geq 1$ form a collision sequence. The collision sequence $(j_1, k) \to \cdots \to (j_t, k)$* ends *in block I_{j_t} if $(j_{t+1}, k) \notin \widehat{I}_{j_{t+1}}$ and $\boldsymbol{g}(j_t, k) \to (j_{t+1}, k)$.*

Collision sequences prevent global expansion. That is, if we have "many" long collision sequences, then there is "not much" opportunity for global expansion.

DEFINITION 5. *Consider* $(j,k_0),(j,k_1)$, *and* (j,k_2) *all from* I_j *so that* $(j,k_1) = \boldsymbol{loc}(j,k_0)$ *and* $(j,k_2) = \boldsymbol{loc}(j,k_1)$, *where*

$$\begin{aligned} &(j,k_0) \notin \widehat{I}_j, \\ &(j,k_1) \in \widehat{I}_j \quad \textit{so } (j,k_1) \textit{ is selected}, \\ &(j,k_2) \notin \widehat{I}_j; \end{aligned}$$

then (j,k_1) *is a* singleton. *A singleton has local expansion of* 1.

Definition 5 is about elements in the same input block I_j. A collision sequence has one or more selected inputs that are all in different input blocks. In fact, a length t collision sequence containing u singletons gives total expansion of at least $u+1$.

The degenerate elements $(j,0)$ for all $j \in \mathbb{Z}_p^+$ do not have global expansion since $\boldsymbol{g}(j,0) = (j,0)$. This means if an adversary can select an element $(j,0)$ to extend a $\boldsymbol{loc}$-contiguous set in $\widehat{I}_j$, then they should do it since it will not give any global expansion. That is, as long as $(j,0)$ would not be a singleton, then selecting it increases the number of elements selected but does not increase any expansion.

Likewise, the degenerate elements $(p-1,k)$ for all $k \in \mathbb{Z}_p^+$ do not have local expansion since $\boldsymbol{loc}(p-1,k) = (p-1,k)$. Therefore, if selecting a $(p-1,k)$ either extends one collision sequence or joins two collision sequences, then an adversary should select it. Selecting such an $(p-1,k)$ will increase the number of selected elements without increasing expansion. In fact, if $(p-1,k)$ joins two collision sequences, then it reduces the overall expansion.

2.3. Double hashing. Hashing with *open addressing* is a storage and search technique on a table T that assumes the number of elements to be stored in the table is at most the table size: $|T|$. Elements or keys are put directly in the table T. No pointers or data structures are used. There is a special element NIL denoting no element in a position it occupies. Given t elements in the table T, the load factor is $\alpha = \frac{t}{|T|}$ and $\alpha < 1$. Generally, the important questions that have arisen for open address hashing are related to the number of probes necessary to find elements in the table.

Consider an open addressing table T of size m and two hash functions h_1 and h_2. Given a key x, determining the $(i+1)$st hash location using *double hashing* is done by

$$h(i,x) = (h_1(x) + i\, h_2(x)) \bmod m.$$

Double hashing is a classical data structure, and discussions of it can be found in [11, 24, 17], for example.

Inserting the element x into the table T is done by first searching for x in T. If T does not contain x, then x can be inserted into T. Likewise, to delete x from T, then it must be determined if x is in T as well as where x is located in T. Therefore, searching for an element x is the focus of studies of double hashing.

The first probe to T is to position $T[h_1(x) \bmod m]$. If $T[h_1(x) \bmod m] =$ NIL, then x is not in T. Otherwise, if x is in $T[h_1(x) \bmod m]$, then double hashing reports where x is: position $h_1(x) \bmod m$ since $i = 0$. If x is not in $T[h_1(x) \bmod m]$, then the next element probed is $T[(h_1(x) + h_2(x)) \bmod m]$ since $i = 1$. If $T[(h_1(x) + h_2(x)) \bmod m]$ is NIL, then x is not in T. Otherwise, if $x = T[(h_1(x) + h_2(x)) \bmod m]$, then the double hashing algorithm is found where x resides and returns the value $(h_1(x) + h_2(x)) \bmod m$. Otherwise, x may still be

in T. Therefore, element $T[(h_1(x) + 2\,h_2(x)) \bmod m]$ is probed, etc. This continues using the function in the $(i+1)$st probe $h(i,x)$ until either x is found or $T[(h_1(x) + i\,h_2(x)) \bmod m]$ is NIL, indicating x is not in T. In summary, the probe sequence is in the following addresses of T:

$$h_1(x),\ (h_1(x) + h_2(x)) \bmod m, \quad (h_1(x) + 2\,h_2(x)) \bmod m,\ \ldots\ .$$

Assume $m > 2$ is prime and h_1 and h_2 are based on ***loc*** and $\boldsymbol{g}$. For a double hash table T, this paper assumes $|T| = m$ as well as the key $x \in \mathbb{Z}_m$.

In the case of this paper's double hashing result, the $\boldsymbol{g}$ edges are the focal point and the ***loc*** edges are not used. In this double hashing scheme, say the pair (j,k) is generated for some key x by h_1 and h_2. That is, start in position k in input block I_j. Hash function h_1 generates the first position j (input block) and hash function h_2 generates the hop-size $k+1$ (how to travel from input block to input block). So the key x is hashed into I_j, starting at local position k. In other words, the first probe is in $T[j]$. If necessary, the second probe is in $T[(j + (k+1)) \bmod m]$. If necessary, the third probe is in $T[(j + 2(k+1)) \bmod m]$, etc.

More precisely, first, a block j_0 and a position k are chosen by h_1 and h_2, respectively. That is, given the key x, compute $j_0 = h_1(x)$ and $k = h_2(x)$. Next, as necessary, the following blocks are computed: $j_1 = \boldsymbol{g}_1(j_0,k)$ and, in general, $j_i = \boldsymbol{g}_1(j_{i-1},k)$, for $i : m-1 \le L \ge i \ge 1$, giving the permutation

$$\langle j_0, \boldsymbol{g}_1(j_0,k), \boldsymbol{g}_1(j_1,k), \ldots, \boldsymbol{g}_1(j_L,k)\rangle.$$

Since m is prime, $\boldsymbol{g}$ sends this permutation exactly once through each of the input blocks $\widehat{I}_0, \widehat{I}_1, \ldots, \widehat{I}_L$, where $L \le m-1$.

The graph G_3, since m a prime, with the functions $\boldsymbol{g}$ represents all permutations used by open addressed double hashing on a table $T[0,\ldots,m-1]$. Of course, $T[j]$ corresponds to I_j.

Double hashing approximates uniform open address hashing [26, 11, 24]. More precisely, Guibas and Szemerédi [14] showed unsuccessful searches using double hashing take asymptotically the same number of probes as idealized uniform hashing does for any fixed load factor α less than about 0.319. For any fixed $\alpha < 1$, see Lueker and Molodowitch [20]. However, as pointed out in Schmidt and Siegel [30], these last results assume ideal randomized functions, whereas [30] utilizes more realistic k-wise independent and uniform functions (where $k = c\log n$ for a suitable constant c).

THEOREM 1 (see [20] and [30]). *Suppose T has load factor of any fixed $\alpha < 1$. The expected number of probes for an unsuccessful search in an open addressing double hashing table is $\frac{1}{1-\alpha} + \epsilon$, where ϵ is a lower-order term.*

Lueker and Molodowitch [20] give the most straightforward method of showing this based on assumed randomized inputs. Schmidt and Siegel [30] give the tightest bound (sharpest bound on ϵ) and the weakest notion of randomness to date. That is, [30] shows the result of Theorem 1 by supplying randomized hash functions, in particular randomized hash functions of degree $c\log n$ for some constant c, giving $c\log n$-wise independent functions.

2.3.1. Strong universal hash functions. Given a graph G_3, where $\alpha = |\widehat{\mathcal{I}}|/|\mathcal{I}|$ for fixed $\alpha : 1 > \alpha > 0$, let $|\widehat{I}_j|/p = \alpha_j$, such that $j \in \mathbb{Z}_p^+$, and each fixed $\alpha_j : 1 > \alpha_j > 0$. This means

$$\begin{aligned}\alpha &= |\widehat{\mathcal{I}}|/|\mathcal{I}| \\ &= \frac{\alpha_0 + \cdots + \alpha_{p-1}}{p}.\end{aligned}$$

Consider any block $\widehat{I}_j$ such that $\mathcal{L}(\widehat{I}_j) \leq 1$. In such blocks the adversary chooses the starting point b_j for the elements of I_j, as well as the total number of elements to select from I_j, expressed here as α_j. More precisely, since $\mathcal{L}(\widehat{I}_j) \leq 1$, the adversary must have chosen the inputs so that $|\boldsymbol{id}(\widehat{I}_j) \cap \boldsymbol{loc}(\widehat{I}_j)| \leq 1$, leaving only the number of elements selected and their starting point to question.

DEFINITION 6 (Carter and Wegman [8]). *The set of function H is strongly universal iff randomly, uniformly, and independently choosing $h \in H$; then for any two different keys x_1 and x_2 and any two values $y_1, y_2 \in \mathbb{Z}_p^+$, it must be that*

$$\mathbf{Pr}[h(x_1) = y_1] = \frac{1}{p} \quad \textit{and}$$
$$\mathbf{Pr}[h(x_1) = y_1 \wedge h(x_2) = y_2] = \frac{1}{p^2}.$$

THEOREM 2 (Carter and Wegman [8]). *The functions*

$$h_{j,b}(x) = jx + b \bmod p \textit{ for all } (j,b) \in \mathbb{Z}_p^+ \times \mathbb{Z}_p^+$$

give the strongly universal set

$$H = \{h_{j,b} \textit{ for all } (j,b) \in \mathbb{Z}_p^+ \times \mathbb{Z}_p^+\}.$$

For all $h \in H$, the range is $\mathbb{Z}_p^+$. Generally, hash functions are expressed as $h_{j,b}(i) \bmod m$, where m is the table size, but here $m = p$, allowing the focus to be entirely on $h_{j,b}(i)$.

2.4. Counting frequencies of selected elements. The basic progression from this subsection to section 3 works as follows. We start with the case where an adversary selects the same number of input elements in each position in all I_j, for $j : j \in \mathbb{Z}_p^+$, while maintaining ***loc***-contiguity of the elements in each $\widehat{I}_j$. Basic bounds on the expansion are developed in this subsection. Subsequent subsections in section 2 incrementally allow an adversary to select any elements they choose as long as they maintain ***loc***-contiguity.

This subsection applies to both universal hashing as well as expansion.

THEOREM 3 (Chor and Goldreich [9]). *Take $h_{j,b} \in H$ uniformly at random; then the associated values $h_{j,b}(i), \ldots, h_{j,b}(0)$, for $p > L \geq i \geq 1$ and $L \geq 2$, are pairwise independent and for all $i \geq 0$ the elements $h_{j,b}(i), \ldots, h_{j,b}(0)$ are uniform in $\mathbb{Z}_p^+$.*

Chor and Goldreich present this result for the sequences of random variables $h_{j,b}(i), \ldots, h_{j,b}(1)$, and it is straightforward that $h_{j,b}(0)$ can be included since $h_{j,b}(0) = b$, which is uniformly and randomly chosen.

Theorem 3 will be applied to $\boldsymbol{g}$ functions between different blocks. Relations *in* the blocks are discussed next. Recall, for each block I_j, the adversary chooses each b_j as well as α_j, and so Theorem 3 does not apply to each block. That is, Theorem 3 assumes the pair $(j,b) \in \mathbb{Z}_p^+ \times \mathbb{Z}_p^+$ is randomly and uniformly chosen.

In contrast to the strongly universal set H of Theorem 2, take $U \subseteq \mathbb{Z}_p^+ \times \mathbb{Z}_p^+$ such that, for all $j \in \mathbb{Z}_p^+$, there is some pair $(j,b) \in U$ and further, if $(j,b_1) \in U$ and $(j,b_2) \in U$, then $b_1 = b_2$. So

$$H' = \{h_{j,b_j}, \text{ for all } (j,b_j) \in U\}, \text{ where } b_j \text{ depends on } j.$$

Note that $|H| = p(p-1)$ and $|H'| = p-1$.

So, in our situation, selecting pairs from H' uniformly at random does not satisfy the hypothesis of this theorem because the adversary chooses each b_j in each pair $(j, b_j) \in H'$.

DEFINITION 7. *Now, for $k \in \mathbb{Z}_p^+$, denote the frequency*

$$n_k = \sum_{j=0}^{p-1} \delta((j,k) \in \widehat{I}_j),$$

where δ is the indicator function, and so $\delta(\mathbf{true}) = 1$ and $\delta(\mathbf{false}) = 0$.

That is, n_k is the frequency of k being selected in all blocks given the adversary's choices of the b_j's and the α_j's.

Note that if $n_k = 0$, then k does not contribute to expansion or lack of expansion. Further, if $n_k = 1$, then k must contribute to global expansion by one.

Aggregating the frequencies gives

$$\alpha = \frac{1}{p^2} \sum_{k=0}^{p-1} n_k.$$

LEMMA 2. *Suppose $n_1 = n_2 = \cdots = n_{p-1}$ and assume $\mathcal{L}(\widehat{I}_j) \leq 1$ for all $j \in \mathbb{Z}_p^+$. Take any randomly and uniformly chosen $(J_1, K) \in \mathbb{Z}_p^+ \times \mathbb{Z}_p^+$, where $(J_2, K) = \boldsymbol{g}(J_1, K)$; then*

$$\mathbf{Pr}[(J_1, K) \in \widehat{I}_{J_1} \wedge (J_2, K) \in \widehat{I}_{J_2}] \leq \frac{n_k^2}{p^2}.$$

Proof. Assume $(J_1, K) \in \mathbb{Z}_p^+ \times \mathbb{Z}_p^+$ is randomly and uniformly chosen. So we are considering the collision sequence,

$$\mathcal{C}_{J_1,K} = (J_1, K) \to (J_2, K),$$

where $(J_2, K) = \boldsymbol{g}(J_1, K)$.

This gives

$$\begin{aligned}
&\mathbf{Pr}[(J_1, K) \in \widehat{I}_{J_1} \wedge (J_2, K) \in \widehat{I}_{J_2}] \\
&= \frac{1}{p^2} \sum_{j=0}^{p-1} \sum_{k=0}^{p-1} \mathbf{Pr}[(j,k) \in \widehat{I}_j \wedge \boldsymbol{g}(j,k) \in \widehat{I}_{\boldsymbol{g}_1(j,k)}] \\
&= \frac{1}{p^3} \sum_{k=0}^{p-1} \sum_{j=0}^{p-1} \delta((j,k) \in \widehat{I}_j)\ \mathbf{Pr}[\boldsymbol{g}(j,k) \in \widehat{I}_{\boldsymbol{g}_1(j,k)} \mid (j,k) \in \widehat{I}_j] \\
&= \frac{1}{p^3} \sum_{k=0}^{p-1} n_k \sum_{j=0}^{p-1} \mathbf{Pr}[\boldsymbol{g}(j,k) \in \widehat{I}_{\boldsymbol{g}_1(j,k)}] \text{ by independence and uniformity of Theorem 3} \\
&= \sum_{k=0}^{p-1} \frac{n_k^2}{p^3} \\
&\leq \frac{n_k^2}{p^2}.
\end{aligned}$$

This completes the proof. □

If all $n_k \leq \alpha p$, then

$$\sum_{k=0}^{p-1} \frac{n_k^2}{p^3} \leq \alpha^2.$$

In other words, let

$$\textbf{Average}[n_k^2] = \sum_{k=0}^{p-1} \frac{n_k^2}{p}.$$

If $n_k \leq \lceil \alpha p \rceil$, for all $k : p > k \geq 0$, then

$$\textbf{Average}[n_k^2] \leq \alpha^2\, p^2$$

so that for uniformly and randomly chosen $(J_1, K) \in \mathbb{Z}_p^+ \times \mathbb{Z}_p^+$ and $(J_2, K) = \boldsymbol{g}(J_1, K)$, then

$$\begin{aligned}\textbf{Pr}[(J_1, K) \in \widehat{I}_{J_1} \wedge (J_2, K) \in \widehat{I}_{J_2}] &\leq \frac{\textbf{Average}[n_k^2]}{p^2} \\ &\leq \alpha^2.\end{aligned}$$

Furthermore, if there is a set

$$\mathbb{K} = \{k_0, \ldots, k_r\}$$

and $|\mathbb{K}| \leq p^\delta$, where $\delta : 1 > \delta \geq 0$, such that

$$n_k > \lceil \alpha(p-1) \rceil \text{ for all } k \in \mathbb{K},$$

then, letting $\overline{\mathbb{K}} = \mathbb{Z}_p^+ - \mathbb{K}$, this gives $n_{k'} \leq \lceil \alpha\, p \rceil$ for all $k' \in \overline{\mathbb{K}}$. This means

$$\begin{aligned}\textbf{Average}[n_k^2] &= \sum_{k \in \mathbb{K}} \frac{n_k^2}{p} + \sum_{k' \in \overline{\mathbb{K}}} \frac{n_{k'}^2}{p} \\ &\leq \frac{p^\delta p^2}{p} + \frac{\alpha^2\, p^2 (p - p^\delta)}{p} \\ &\leq p^\delta p + \alpha^2\, p^2.\end{aligned}$$

Therefore, if $|\mathbb{K}| \leq p^\delta$ for any $\delta : 1 > \delta \geq 0$, and for uniformly and randomly chosen $(J_1, K) \in \mathbb{Z}_p^+ \times \mathbb{Z}_p^+$ and $(J_2, K) = \boldsymbol{g}(J_1, K)$, then

$$\textbf{Pr}[(J_1, K) \in \widehat{I}_{J_1} \wedge (J_2, K) \in \widehat{I}_{J_2}] \leq \alpha^2 + O(p^{\delta - 1}) \text{ for } \delta : 1 > \delta \geq 0.$$

2.5. Bounding Average$[n_k^2]$ when $n_k > \lceil \alpha\, p \rceil$ for $k \in \mathbb{Z}_p^+$. Consider the maximal subset $\mathbb{K} \subseteq \mathbb{Z}_p^+$, where all $k' \in \mathbb{K}$ are such that the frequencies $n_{k'} > \lceil \alpha\, p \rceil$ when $\mathcal{L}(\widehat{\mathcal{I}}) < \frac{3}{16} |\widehat{\mathcal{I}}|$.

Start with the case $\mathcal{L}(\widehat{I}_j) \leq 1$ for all $j \in \mathbb{Z}_p^+$. This subsection shows for all $k' \in \mathbb{K} \subseteq \mathbb{Z}_p^+$ such that $n_{k'} > \lceil \alpha\, p \rceil$ there can be a total of at most $p \lceil 1/\alpha \rceil$ total selected elements in all collision sequences of length more than $\lceil \alpha\, p \rceil$ for all $k \in \mathbb{Z}_p^+$.

Definition 8. *Let $\alpha = |\widehat{\mathcal{I}}|/|\mathcal{I}|$. If there is some $k \in \mathbb{Z}_p^+$ so that k's frequency n_k is such that*

$$n_k > \lceil \alpha\, p \rceil,$$

then there are $n_k - \lceil \alpha\, p \rceil$ excess *elements selected in the kth position of $\mathcal{I}$.*

2.5.1. The case with up to 1 excess element selected. Here k_0 denotes a single excess element.

DEFINITION 9. *Let $n_{k,L}$ denote any n_k when all $j_r \in \{j_0, \ldots, j_{t-1}\}$ are all such that $|\widehat{I}_{j_r}| = L$.*

Further, by Definition 9 it must be that $\alpha' = L/p$ and all n_k's throughout the rest of this section are associated only with the blocks indexed by $\{j_0, \ldots, j_{t-1}\}$.

For the next lemma recall h_{j_i} is a hash function representing the ***loc*** functions corresponding to $\widehat{I}_{j_i}$.

LEMMA 3. *Let $\mathcal{L}(\widehat{I}_j) \leq 1$ for all $j \in \mathbb{Z}_p^+$. Assume k_0 is selected in all of $I_{j_0}, \ldots, I_{j_{t-1}}$ and $|\widehat{I}_{j_r}| = L \leq p$, for $j_r \in \{j_0, \ldots, j_{t-1}\}$, $t \geq L \geq 2$, with $h_{j_0}(v) = \cdots = h_{j_{t-1}}(v) = k_0$ for some $v \in \mathbb{Z}_p^+$; then $n_k \leq \lceil(\alpha' - \frac{1}{p})\ p\rceil$ for all $k \neq k_0$ and $\alpha' = L/p$.*

Proof. The proof is by induction on the size of $|\widehat{I}_{j_r}| = L$. For the moment, assume $v = 0$ for the induction; this will be generalized after the induction is complete.

Basis. Consider the case where $|\widehat{I}_{j_r}| = 2 = L$ for all $j_r \in \{j_0, \ldots, j_{t-1}\}$, where $t \geq L$, and $h_{j_0}(0) = \cdots = h_{j_{t-1}}(0) = k_0$. Then all elements $u_r = (j_r + 1 + k_0) \bmod p$, for all $r : t > r \geq 0$, are distinct since $\{j_0, \ldots, j_{t-1}\}$ are distinct and $t \geq 2$. That is, if $u_{r_x} = u_{r_y}$, then

$$(j_x + 1 + k_0) \bmod p = (j_y + 1 + k_0) \bmod p,$$

and so $j_x = j_y \bmod p$, and it must be that $j_x < p$ and $j_y < p$, and thus $x = y$, a contradiction. So $n_k = 1$, for all $k \neq k_0$, and $\alpha' = \frac{2}{p}$. Further, all $k \neq k_0$ are such that $n_k \leq \lceil(\alpha' - \frac{1}{p})\, t\rceil$, completing the basis.

Inductive hypothesis. Suppose $|\widehat{I}_{j_r}| = L \geq 2$, where $j_r \in \{j_0, \ldots, j_{t-1}\}$ and $t \geq L$. Then $n_{k,L} \leq \lceil(\alpha' - \frac{1}{p})\ t\rceil$ for all $k \neq k_0$ and $\alpha' = \frac{L}{p}$.

Inductive step. Suppose $|\widehat{I}_{j_r}| = L+1$, where $j_r \in \{j_0, \ldots, j_{t-1}\}$ and $t \geq L+1 \geq 3$, where α' is associated with $n_{k,L+1}$. Then by the inductive hypothesis the first L elements in each block share the property that all but one of the $n_{k,L}$'s are such that $n_{k,L} \leq \lceil(\alpha' - \frac{1}{p})\ t\rceil$. Adding one element to each block and each in a unique position gives $n_{k,L+1} \leq \lceil(\alpha' + \frac{1}{p} - \frac{1}{p})\ t\rceil$. Each element is put in a unique position since $t \geq L$, and no two elements among $u_r = (L + 1)(j_r + 1) + k_0 \bmod p$, for $r \in \{t - 1, \ldots, 0\}$, are the same: If $j_x \neq j_y$, where

$$((L + 1)(j_x + 1) + k_0) = ((L + 1)(j_y + 1) + k_0) \bmod p,$$

then since each element of a field (mod-p) has a multiplicative and additive inverse, we must have $j_x = j_y \bmod p$, a contradiction since $j_x < p$ and $j_y < p$. Therefore, applying the inductive hypothesis completes the induction.

Now we show this lemma holds for any $v \in \mathbb{Z}_p^+$, where $h_{j_0}(v) = \cdots = h_{j_{t-1}}(v) = k_0$. Consider $|\widehat{I}_{j_r}| = L$ for all $j_r \in \{j_0, \ldots, j_{t-1}\}$ and $t \geq L$, and given some $v \neq 0$, then break the problem into two cases: The first case consists of all elements $h_{j_0}(i), \ldots, h_{j_{t-1}}(i)$ for $i : L > i \geq v$. The second case consists of all elements $h_{j_0}(i'), \ldots, h_{j_{t-1}}(i')$ for $i' : v \geq i' \geq 0$. (Note that these cases overlap since they both have k_0 in common.)

Now, treating these cases separately, apply the induction above with $\alpha_1 = \frac{L-v}{p}$ to the $L - v$ elements of $h_{j_0}(i), \ldots, h_{j_{t-1}}(i)$ for $i : L > i \geq v$.

Likewise, for each $h_{j_0}(i'), \ldots, h_{j_{t-1}}(i')$, where $i' : v \geq i' \geq 0$, apply the induction above to the v elements. Here $\alpha_2 = \frac{v+1}{p}$.

Now $\alpha' = \alpha_1 + \alpha_2 - \frac{1}{p}$ since k_0 was counted twice. With this in mind, take the following inequalities, where $k \neq k_0$:

$$\begin{aligned} n_k &\leq \left\lceil \left(\alpha_1 - \frac{1}{p}\right) t \right\rceil + \left\lceil \left(\alpha_2 - \frac{1}{p}\right) t \right\rceil \\ &\leq \left\lceil \left(\alpha_1 + \alpha_2 - \frac{2}{p}\right) p \right\rceil \\ &\leq \left\lceil \left(\alpha' + \frac{1}{p} - \frac{2}{p}\right) p \right\rceil \\ &\leq \left\lceil \left(\alpha' - \frac{1}{p}\right) p \right\rceil, \end{aligned}$$

completing the proof. □

With a little work, Lemma 3 generalizes to Theorem 4. Assume $\mathcal{L}(\widehat{I}_j) \leq 1$, for all $j \in \mathbb{Z}_p^+$ and k_0, is selected in each block $I_0, \ldots, I_{p-2}$. Let L_u be the number of elements selected going "above" and including k_0. Similarly, let L_d be the number of elements selected going "down" from k_0 but not including k_0. (If $k_0 = h(i)$, then $h(i+c)$ is "above" for any integer c, where $i + c < p$, and $h(i-c)$ is "down," where $i - c \geq 0$.) The induction is about the same; the only difference is the proof of Theorem 4 assumes the relation

$$n_k \leq \lceil (\alpha_u + \alpha_d)\, t \rceil$$

holds before the inductive step. More precisely, suppose $t = \lceil \frac{p}{c} \rceil$ for some integer c, and by definition $\alpha_u + \alpha_d = \frac{L_u + L_d}{p}$, meaning

$$\left(\frac{L_u + L_d}{p}\right) t \leq \frac{L_u + L_d}{c}.$$

So there are a total of $L_u + L_d$ elements selected per input block, and $\frac{1}{c}$ bounds the percent of blocks under consideration. The inductive hypothesis says $\lceil \frac{L_u + L_d}{c} \rceil$ is an upper bound on the number of elements for each $k \neq k_0$.

Now consider adding one element to each block going "up" and one element to each block going "down." That is, increase α_u to $\alpha_u + \frac{1}{p}$ and increase α_d to $\alpha_d + \frac{1}{p}$, but at the same time none of the new "up" elements collide with each other and none of the new elements going "down" collide with each other. That is, for all $r \in \{0, 1, \ldots, t-1\}$,

$$u_r = (L_u + 1)(j_r + 1) \bmod p$$

are all different by the uniqueness of multiplicative inverses in $\mathbb{Z}_p^+ - \{0\}$. Likewise, for all $r \in \{0, 1, \ldots, t-1\}$,

$$d_r = (L_d + 1)(j_r + 1) \bmod p$$

are all different by the uniqueness of multiplicative inverses in $\mathbb{Z}_p^+ - \{0\}$. Furthermore, by the uniqueness of multiplicative inverses in $\mathbb{Z}_p^+ - \{0\}$, for each d_i there can be at most one u_j so that $d_i = u_j$, where $i, j \in \{0, 1, \ldots, t-1\}$. Consider increasing L_u to $L_u + 1$ and increasing L_d to $L_d + 1$, and assume $t = \lfloor \frac{p}{c} \rfloor$, for some integer c, giving

$$\left(\frac{L_u + L_d + 2}{p}\right) t \leq \frac{L_u + L_d + 2}{c},$$

which is the number of selected elements per input block multiplied by the percent of the input blocks. Therefore, $n_k \leq \lceil (\alpha_u + \alpha_d)\, t + \frac{2\,t}{p} \rceil$, giving, for all $k \neq k_0$,

$$n_k \leq \left\lceil \left(\alpha_u + \alpha_d + \frac{2}{p} \right) t \right\rceil, \tag{1}$$

which clearly holds for $t > p/2$ and $(\alpha_u + \alpha_d)\, t$ bounded by an integer. In the case where $t < p/2$, then since $L_u + L_d + 2$ elements were selected per input block and their ***loc***-continuity gives for each $k \neq k_0$, by the inductive hypothesis $c\, n_k \leq L_u + L_d$ and by the uniqueness of multiplicative inverses, n_k can increase no more than 2 when both L_u and L_d are increased by 1 each. That is, now $c\, n_k \leq L_u + L_d + 2$ holds, completing the inductive step. This gives the next theorem.

THEOREM 4. *Let $\mathcal{L}(\widehat{I}_j) \leq 1$ for all $j \in \mathbb{Z}_p^+$. Assume k_0 is selected in all of $I_{j_0}, \ldots, I_{j_{t-1}}$ and $|\widehat{I}_{j_r}| = L \leq p$, for $j_r \in \{j_0, \ldots, j_{t-1}\}$, $t \geq L \geq 2$, with $h_{j_0}(v) = \cdots = h_{j_{t-1}}(v) = k_0$ for some $v \in \mathbb{Z}_p^+$; then $n_k \leq \lceil \alpha'\, t \rceil$ for all $k \neq k_0$ and $\alpha' = L/p$.*

The next lemma allows any of t blocks to have any number of elements L selected as long as $t \geq L$. That is, $|\widehat{I}_{j_r}| \leq L$ for all $j_r \in \{j_0, \ldots, j_{t-1}\}$ and $t \geq L$.

LEMMA 4. *Let $\mathcal{L}(\widehat{I}_{j_r}) \leq 1$ for all $j_r \in \{j_0, \ldots, j_{t-1}\}$. Assume k_0 is selected in all of $I_{j_0}, \ldots, I_{j_{t-1}}$ and $|\widehat{I}_{j_r}| \leq L \leq p$ for $j_r \in \{j_0, \ldots, j_{t-1}\}$ and $t \geq L \geq 2$ while $h_{j_0}(v) = \cdots = h_{j_{t-1}}(v) = k_0$ for some $v \in \mathbb{Z}_p^+$; then $n_k \leq \lceil \alpha'\, t \rceil$ for all $k \neq k_0$ and $\alpha' = (|\widehat{I}_{j_0}| + \cdots + |\widehat{I}_{j_{t-1}}|)/(t\, p)$.*

Proof. Consider two sets T_1 and T_2 with $s : t \geq s \geq 0$, so that

$$T_1 = \{\ \widehat{I}_{j_0}, \ldots, \widehat{I}_{j_{s-1}}\ \}, \quad \text{where } \ \alpha_1\, p = |\widehat{I}_{j_k}| \text{ for all } k : s - 1 \geq k \geq 0,$$

and

$$T_2 = \{\ \widehat{I}_{j_s}, \ldots, \widehat{I}_{j_{t-1}}\ \}, \quad \text{where } \ \alpha_2\, p = |\widehat{I}_{j_k}| \text{ for all } k : t - 1 \geq k \geq s.$$

Therefore, there is a total of $\alpha'\, p\, t$ selected elements in all of the blocks

$$\{\ \widehat{I}_{j_0}, \ldots, \widehat{I}_{j_{t-1}}\ \},$$

and so

$$\alpha'\, p\, t = \alpha_1\, p\, s + \alpha_2\, p\, (t - s).$$

That is,

$$\alpha'\, t = \alpha_1\, s + \alpha_2\, (t - s).$$

Without loss, assume $\alpha_1 < \alpha_2$ and T_1 represents the first s input blocks. Now, applying Theorem 4 to all t input blocks considering only $\alpha_{\min} = \min\{\ \alpha_1, \alpha_2\ \}$, it must be that

$$n_k^0 \leq \lceil \alpha_{\min}\ \ t \rceil$$

for all $k \neq k_0$, and each n_k^0 is computed restricting k to the $\alpha_{\min}\ t$ elements of each of all t input blocks. Note that Theorem 4 applies to the first $\alpha_{\min} t$ ***loc***-contiguous elements from each block since $t \geq L$. Without loss, assume $\alpha_{\min} t$ is an integer.

Now, letting $\alpha_{\max} = \max\{\ \alpha_1, \alpha_2\ \}$, then since $t \geq L$ and assuming $t = \lceil \frac{p}{c} \rceil$ for some integer c, then applying Theorem 4

$$n_k^1 \leq \lceil (\alpha_{\max} - \alpha_{\min})\ t \rceil$$

for all $k \neq k_0$. But, not considering the $\lceil \alpha_{\min} t \rceil$ elements, it must be that

$$n_k^1 \leq \lceil (\alpha_{\max} - \alpha_{\min})\ (t-s) \rceil .$$

Without loss, assume that $(\alpha_{\max} - \alpha_{\min})\ (t-s)$ and $\alpha_{\min}\ t$ are integers. This means, for $k \neq k_0$, and since the elements represented by $\alpha_{\max}$ and $\alpha_{\min}$ are ***loc***-contiguous,

$$\begin{aligned} n_k^0 + n_k^1 &\leq \lceil \alpha_{\min} t + (\alpha_{\max} - \alpha_{\min})\ (t-s) \rceil \\ &\leq \lceil \alpha_{\max} (t-s) - \alpha_{\min} (t-s) + \alpha_{\min} t \rceil \\ &\leq \lceil \alpha_{\max} (t-s) + \alpha_{\min} s \rceil \\ &\leq \lceil \alpha_1 s + \alpha_2 (t-s) \rceil \\ &\leq \lceil \alpha' t \rceil , \end{aligned}$$

since $\alpha_{\max} = \alpha_2$ and $\alpha_{\min} = \alpha_1$ and $\alpha' t = \alpha_1 s + \alpha_2 (t-s)$, while at the same time $n_k^0 + n_k^1 > \lceil (\alpha_1 + \alpha_2)(t-s) \rceil$ only for $k = k_0$; the proof is completed by induction on the number of sets of blocks, each set containing blocks with the same number of elements selected. □

Now consider combining ***loc***-contiguous collision sequences as described by Lemma 4. In particular, now look at bounding the length of all collision sequences in G_3 by combining different ***loc***-contiguous subblocks.

The next definition generalizes Definition 2.

DEFINITION 10. *Take $j \in \mathbb{Z}_p^+$. The set $U_{j,s}$ is a* maximal ***loc***-contiguous subblock *of $\widehat{I}_j$ if $U_{j,s} \subset \widehat{I}_j$. And if $|U_{j,s}| \geq 2$, while $\mathcal{L}(U_{j,s}) \leq 1$, and for any $U' \subset \widehat{I}_j : U_{j,s} \subset U'$ and $U_{j,s} \neq U'$, then $\mathcal{L}(U') > 1$. Further, if $U_{j,0} = I_j$ (so $|U_{j,0}| = p$), then $U_{j,0}$ is the only* maximal ***loc***-contiguous subblock *of $\widehat{I}_j$.*

Note that a maximal ***loc***-contiguous subblock $U_{j,s}$ must contain at least two elements; otherwise, it is not ***loc***-contiguous. Further, a block $\widehat{I}_j$ may have many maximal ***loc***-contiguous subblocks. Allow maximal ***loc***-contiguous subblocks to be empty. This means maximal ***loc***-contiguous subblocks cannot consist of a single ***loc***-contiguous element.

DEFINITION 11. *Consider the collision sequence $\mathcal{C}_1$ made up of selected elements from the max* ***loc****-contiguous subblocks $U_{0,0}, \ldots, U_{p-1,0}$. Another collision sequence $\mathcal{C}_2$* overlaps *with $\mathcal{C}_1$ iff $\mathcal{C}_2$ consists of elements from at least one of the same max* ***loc****-contiguous subblocks $U_{0,0}, \ldots, U_{p-1,0}$.*

So now remove from consideration all subblocks associated with any collision sequence in c_0, i.e., $U_{0,0}, \ldots, U_{p-1,0}$. Now with the remaining elements, put the next largest collision sequences in a set c_1. Any collision sequence $\mathcal{C}_r \in c_1$ is associated with the sets of max ***loc***-contiguous sequences $U_{0,r}, \ldots, U_{p-1,r}$. By Lemma 4 and Theorem 4 all elements of any other collision sequence can share no more than $\lceil \alpha_1 p \rceil$ elements with $U_{0,r}, \ldots, U_{p-1,r}$, where $\alpha_1 = (|U_{0,r}| + \cdots + |U_{p-1,r}|)/p^2$.

This argument extends to all collision sequences larger than $\lceil \alpha p \rceil$.

LEMMA 5. *Let $\mathbb{K}$ be the set of indices of all s collision sequences larger than $\lceil \alpha p \rceil$. Suppose the adversary selects no singletons and $\alpha_i = (|U_{0,i}| + \cdots + |U_{p-1,i}|)/p^2$, where n_k^i is n_k* restricted *to the jth set of max* ***loc****-contiguous subblocks $U_{j,i}$, for i, where $i : s \geq i \geq 0$ and all $j : p-1 \geq j \geq 0$. So for any $k \in \mathbb{Z}_p^+ - \mathbb{K}$, then*

$$n_k^0 + \cdots + n_k^s \leq \left\lceil \left(\alpha_0 - \frac{1}{p} \right) p \right\rceil + \left\lceil \left(\alpha_1 - \frac{1}{p} \right) p \right\rceil + \cdots + \left\lceil \left(\alpha_s - \frac{1}{p} \right) p \right\rceil .$$

Proof. Without loss, let $\mathcal{C}_0, \mathcal{C}_1, \ldots, \mathcal{C}_s$ be collision sequences of lengths larger than $\lceil \alpha\, p \rceil$. Assume these are listed largest ($\mathcal{C}_0$) to smallest ($\mathcal{C}_s$). No two collision sequences $\mathcal{C}_i$ and $\mathcal{C}_j$, for $i \neq j$, are made up of elements from the same max ***loc***-contiguous subblocks $U_{j_0,i'}, \ldots, U_{j_{k-1},i'}$ for some i', by Lemma 4 and Theorem 4. In the case where two collision sequences share a max ***loc***-contiguous subblock, then this max ***loc***-contiguous subblock can be cut into two ***loc***-contiguous subblocks. Using this fact, starting with the largest collision sequences first, each collision sequence is uniquely associated with a set of $p-1$ ***loc***-contiguous subblocks, one for each input block I_j, for $j \in \mathbb{Z}_p^+$. (Some of these ***loc***-contiguous subblocks may be empty.)

This means each collision sequence is associated with one ***loc***-contiguous subblock from each input block,

$$\langle U_{0,0}, \ldots, U_{p-1,0} \rangle\,, \quad \langle U_{0,1}, \ldots, U_{p-1,1} \rangle\,, \ldots, \langle U_{0,s}, \ldots, U_{p-1,s} \rangle\,,$$

and in some cases $U_{j,i} = \{\emptyset\}$.

Let $\mathbb{K} = \{k_0, \ldots, k_s\} \subset \mathbb{Z}_p^+$ be such that $k \in \mathbb{K}$ means $n_k \geq \lceil \alpha\, p \rceil$, and now take it as $n_k = p-1$. If $k \in \mathbb{Z}_p^+ - \mathbb{K}$ and n_k^i is restricted to $U_{0,i}, \ldots, U_{p-1,i}$, where $\alpha_i = (|U_{0,i}| + \cdots + |U_{p-1,i}|)/p(p-1)$, then

$$n_k^i \leq \left\lceil \left(\alpha_i - \frac{1}{p} \right) p \right\rceil,$$

which gives the lemma. □

The next lemma deals with the case when the number of elements in each block is larger than the total number of such blocks, or $L > t$.

LEMMA 6. *Let $\mathcal{L}(\widehat{I}_j) \leq 1$ for all $j \in \mathbb{Z}_p^+$. Given t blocks $I_{j_0}, \ldots, I_{j_{t-1}}$, where $|\widehat{I}_{j_r}| \leq L \leq p$, for $j_r \in \{j_0, \ldots, j_{t-1}\}$ so that $\alpha' = (|\widehat{I}_{j_0}| + \cdots + |\widehat{I}_{j_{t-1}}|)/(t\ p)$, and letting $S = |\boldsymbol{id}_2(\widehat{I}_{j_0}) \cup \cdots \cup \boldsymbol{id}_2(\widehat{I}_{j_{t-1}})|$, now let*

$$T = \left\lceil \frac{S}{t} \right\rceil;$$

then at most T of the n_k's are such that $n_k > \lceil \alpha' t \rceil$.

Proof. Take the L elements in each block and consider them in $T = \lceil \frac{S}{t} \rceil$ ***loc***-contiguous sets of inputs of size t each, in every block. If $T = 1$, then we are done. Next, consider each set of t selected ***loc***-contiguous input elements among the t blocks; then assume there is an input set $\{i_0, \ldots, i_{t-1}\}$, so that $h_{j_0}(i_0) = \cdots = h_{j_{t-1}}(i_{t-1})$. If not, then consider the largest such input set for each of the t selected ***loc***-contiguous sets of elements. By Theorem 4 and since there are up to t elements in each set of ***loc***-contiguous elements, then each set of ***loc***-contiguous blocks alone has at most one $n_{k'}$ such that $n_{k'} > \lceil \alpha_s\, t \rceil$, for $\alpha_s = \alpha'/T$, for some k'.

This means, among all T size t ***loc***-contiguous sets among the t input blocks, there will be at most T elements $n_{k_i'} > \lceil \alpha_s\, t \rceil$ for $\alpha_s = \frac{t}{L}\alpha'$ and $i : T-1 \geq i \geq 0$. Let

$$\mathbb{K} = \{\ k_0, \ldots, k_{T-1}\ \}$$

be the set of n_{k_i} elements so that $n_{k_i} > \lceil \alpha_s\, t \rceil$.

Let n_k^ℓ denote n_k restricted to the ℓth set of ***loc***-contiguous blocks. In other words, n_k^ℓ is the number of times k occurs in the ℓth set of ***loc***-contiguous blocks.

Next, without loss, suppose that $\alpha_s\, t = \frac{L\, t}{p}$ is an integer, and considering all ***loc***-contiguous sets at the same time gives, for any $k_i \notin \mathbb{K}$,

$$\begin{aligned} n^0_{k_i} + \cdots + n^{T-1}_{k_i} &\leq \lceil (\alpha_s + \cdots + \alpha_s)\, t \rceil, \text{ where there are } T \text{ total } \alpha_s \\ &\leq \lceil \alpha'\, t \rceil, \end{aligned}$$

by Lemma 5.

This completes the proof. □

Suppose there is a collision sequence of length t made of one selected element from each of $\widehat{I}_{j_0}, \ldots, \widehat{I}_{j_{t-1}}$. If $t \geq \lceil \alpha\, p \rceil$, then Lemma 6 indicates that

$$\begin{aligned} T &\leq \left\lceil \frac{S}{\alpha\, p} \right\rceil \\ &\leq \left\lceil \frac{1}{\alpha} \right\rceil, \end{aligned}$$

since $p \geq S$, and where $S = |\boldsymbol{id}_2(\widehat{I}_{j_0}) \cup \cdots \cup \boldsymbol{id}_2(\widehat{I}_{j_{t-1}})|$. That is, suppose this occurs in a set of t input blocks where $S > t$ and the largest collision sequences are of length at least $\lceil \alpha\, p \rceil$. The only selected elements in excess of $\lceil \alpha\, p \rceil$ in $T = \lceil \frac{S}{\alpha} \rceil$ "large" collision sequences can be larger than $\lceil \alpha\, p \rceil$. Since $T \leq \lceil 1/\alpha \rceil$, if all of these T "large" collision sequences consist of p elements each, there is a total of at most

$$\frac{p - \lceil \alpha\, p \rceil}{\alpha} \leq \frac{p}{\alpha}$$

excess elements out of a total of p^2 possible elements and $\alpha\, p^2$ selected elements. That is, the uniform random probability of selecting an element that extends a collision sequence to one of these excess elements is at most $\frac{1}{\alpha\, p}$.

It is also possible that the adversary chooses $t > \lceil \alpha\, p \rceil$ blocks that have more elements selected in each block than αp^2, where $\alpha = |\widehat{\mathcal{I}}|/|\mathcal{I}|$. Take the case where $\alpha' > \alpha$, where $\alpha = |\widehat{\mathcal{I}}|/|\mathcal{I}|$, and take $\alpha = \frac{L}{p}$ for appropriate L, but say $\alpha' \leq \frac{L+d}{p}$, for some integer d, for all $\widehat{I}_{j_0}, \ldots, \widehat{I}_{j_{t-1}}$. This means

$$T = \left\lceil \frac{L + d}{\lceil \alpha\, p \rceil} \right\rceil,$$

but since $\alpha = \frac{L}{p}$, it must be that $\alpha\, p = L$, and therefore

$$T = 1 + \left\lceil \frac{d}{\lceil \alpha\, p \rceil} \right\rceil.$$

Assuming $d > \lceil \alpha\, p \rceil$, then $T \leq 1 + \lceil \frac{1}{\alpha} \rceil$, since $d < p$, and thus say $d = \frac{p}{c}$ for some number $c \geq 1$; then

$$\begin{aligned} \left\lceil \frac{d}{\lceil \alpha\, p \rceil} \right\rceil &= \left\lceil \frac{p}{c \lceil \alpha\, p \rceil} \right\rceil \\ &\leq \frac{1}{c\, \alpha} \\ &\leq \frac{1}{\alpha} \end{aligned}$$

since $1 > \alpha$ and $c \geq 1$. So now we discard any case where $L > t$ by this discussion and Lemma 6.

2.5.2. When more than one excess element is selected in $\widehat{\mathcal{I}}$. The next theorem assumes no more than one ***loc***-contiguous subblock is selected per block I_j for $j \in \mathbb{Z}_p^+$. Therefore, $\mathcal{L}(\widehat{I}_j) \leq 1$ for all $j \in \mathbb{Z}_p^+$, which means at most p total elements of all of the n_k's are such that $n_k > \lceil(\alpha - \frac{1}{p})\, p\rceil$.

DEFINITION 12. *Given the frequency n_k, then $B[n_k] \subseteq \mathbb{Z}_p^+$ is the set of block indices that have element k selected in each of them. That is, $u \in B[n_k]$ iff $\delta((u,k) \in \widehat{I}_u)$.*

THEOREM 5. *Suppose $\mathcal{L}(\widehat{I}_j) \leq 1$ for all blocks $j \in \mathbb{Z}_p^+$. Then the total number of excess elements in G_3 is at most $p\lceil 1/\alpha\rceil$.*

Proof. Consider the frequencies, $n_{k_0} \geq n_{k_1} \geq \cdots \geq n_{k_s} > \lceil\alpha\ p\rceil$. By Lemma 5, this proof must consider only the elements $\{k_0, \ldots, k_s\} \in \mathbb{Z}_p^+$. Further, by Lemma 6, at most $T = \lceil 1/\alpha\rceil$ frequencies are such that $|B[n_{k_i}] \cap B[n_{k_j}]| \geq \lceil\alpha p\rceil$ for $k_i \neq k_j$.

Now, if all $k, k' \in \{k_0, \ldots, k_s\}$ are such that

$$B[n_k] \cap B[n_{k'}] \neq \emptyset,$$

then clearly

$$\sum_{i=0}^{s} n_{k_i} \leq p,$$

which would complete the proof.

Furthermore, if any subset $\{u_0, \ldots, u_t\} \subseteq \{k_0, \ldots, k_s\}$ is such that for all $k_i \in \{k_0, \ldots, k_s\}$ and for all $u_i \in \{u_0, \ldots, u_t\}$,

$$B[n_{u_i}] \cap B[n_{k_i}] = \emptyset,$$

then we need only consider the set

$$\mathbb{K} = \{k_0, \ldots, k_s\} - \{u_0, \ldots, u_t\}.$$

Given two distinct collision sequences, then by Lemma 4 these collision sequences can share no more than $\lceil\alpha\, p\rceil$ elements as long as $\mathcal{L}(\widehat{I}_j) \leq 1$.

This means

$$|B[n_{k_0}] \cap (B[n_{k_1}] \cup \cdots \cup B[n_{k_s}])| \leq \lceil\alpha p\rceil,$$

and thus removing the block indices $B[n_{k_0}]$ and by applying Lemma 4 again gives

$$|B[n_{k_1}] \cap (B[n_{k_2}] \cup \cdots \cup B[n_{k_s}])| \leq \lceil\alpha p\rceil.$$

The proof is completed by induction on the remaining blocks $B[n_{k_2}], \ldots, B[n_{k_s}]$. ∎

That is, if $\mathcal{L}(\widehat{I}_j) \leq 1$, for all $j \in \mathbb{Z}_p^+$, then the total number of excess selected elements is $\lceil 1/\alpha\rceil p$. This means the probability of uniformly and randomly selecting one of these up to p *excess* elements is at most $\lceil\frac{p}{\alpha\, p^2}\rceil = \lceil\frac{1}{\alpha\, p}\rceil$.

Next, this is generalized to the case where $\mathcal{L}(\widehat{\mathcal{I}}) \leq \frac{3}{16}|\widehat{\mathcal{I}}|$. In this case, this subsection concludes by showing for all $k \in \mathbb{Z}_p^+$ that $\mathbf{Average}[n_k^2]$ is bounded so that randomly, independently, and uniformly choosing (J_1, K) from $\mathbb{Z}_p^+ \times \mathbb{Z}_p^+$, where $(J_2, K) = \boldsymbol{g}(J_1, K)$, gives

$$\mathbf{Pr}[(J_1, K) \in \widehat{I}_{J_1} \wedge (J_2, K) \in \widehat{I}_{J_2}] \leq \alpha^2 + \epsilon,$$

where $\epsilon = O(\frac{1}{p})$.

As before, start assuming $\mathcal{L}(\widehat{I}_j) \leq 1$ for all blocks $j \in \mathbb{Z}_p^+$. Take the relations,

$$n_{k_0} \geq n_{k_1} \geq \cdots \geq n_{k_s} > \lceil \alpha\, p \rceil,$$

so that $n_k \leq \lceil \alpha\, p \rceil$ for all $k \in \mathbb{Z}_p^+ - \mathbb{K}$, where $\mathbb{K} = \{\ k_0, \ldots, k_s\ \}$.

THEOREM 6. *Suppose $\mathcal{L}(\widehat{I}_j) \leq 1$ for all $j \in \mathbb{Z}_p^+$ and $\alpha = |\widehat{\mathcal{I}}|/|\mathcal{I}|$. Take any randomly and uniformly chosen $(J_1, K) \in \mathbb{Z}_p^+ \times \mathbb{Z}_p^+$, where $(J_2, K) = \boldsymbol{g}(J_1, K)$; then*

$$\mathbf{Pr}[(J_1, K) \in \widehat{I}_{J_1} \wedge (J_2, K) \in \widehat{I}_{J_2}] \leq \alpha^2 + \epsilon,$$

where ϵ is of lower-order terms.

Proof. If there is at most one frequency $n_{k_0} > \lceil \alpha\, p \rceil$, then for all of the $p^2 - p$ or more *non*excess elements in G_3, it must be that

$$\mathbf{Pr}[(J_1, K) \in \widehat{I}_{J_1} \wedge (J_2, K) \in \widehat{I}_{J_2}] \leq \alpha^2$$

holds by Lemma 2.

Consider the frequencies $n_{k_0} \geq n_{k_1} \geq \cdots \geq n_{k_s} > \lceil \alpha\, p \rceil$, where $s \geq 1$. The case of the up to p elements in $n_{k_0}, \ldots, n_{k_s}$, the probability of them extending a one-element collision sequence with excess elements by Theorem 5, is

$$\left\lceil \frac{p}{\alpha p^2} \right\rceil = \left\lceil \frac{1}{\alpha p} \right\rceil.$$

Thus, for all the elements in G_3 it must be that

$$\mathbf{Pr}[(J_1, K) \in \widehat{I}_{J_1} \wedge (J_2, K) \in \widehat{I}_{J_2}] \leq \alpha^2 + O\left(\frac{1}{p}\right)$$

holds.

This completes the proof. □

If $\mathcal{L}(\widehat{I}_j) > 1$ for some $j \in \mathbb{Z}_p^+$, then there may be many collision sequences that must be considered.

Start with G_3 where the adversary has selected $|\widehat{\mathcal{I}}|$ input elements. Then consider any set of at most $\frac{3}{16}|\widehat{\mathcal{I}}|$ sets of associated ***loc***-contiguous elements, where each associated set contains a common collision sequence. By applying Theorem 6 to each collision sequence created by increasing local expansion to decrease global expansion (by extending or joining collision sequences) gives the next theorem. Note that each extended collision sequence has some associated $\alpha_i < \alpha = |\widehat{\mathcal{I}}|/|\mathcal{I}|$ and $\alpha_1 + \cdots + \alpha_u = \alpha$, and so $\alpha_1^2 + \cdots + \alpha_u^2 \leq \alpha^2$. This is because

$$\begin{aligned} \alpha_1^2 + \cdots + \alpha_u^2 &\leq (\alpha_1 + \cdots + \alpha_u)^2 \\ &\leq \alpha^2. \end{aligned}$$

THEOREM 7. *Let $\alpha = |\widehat{\mathcal{I}}|/|\mathcal{I}|$. Suppose $\mathcal{L}(\widehat{\mathcal{I}}) < \frac{3}{16}|\widehat{\mathcal{I}}|$ and all selected elements are in a **loc**-contiguous subblocks and there are no singletons. Take any randomly and uniformly chosen $(J_1, K) \in \mathbb{Z}_p^+ \times \mathbb{Z}_p^+$, where $(J_2, K) = \boldsymbol{g}(J_1, K)$; then*

$$\mathbf{Pr}[(J_1, K) \in \widehat{I}_{J_1} \wedge (J_2, K) \in \widehat{I}_{J_2}] \leq \alpha^2 + \epsilon,$$

where ϵ is of lower-order terms.

3. Variations on hashing. Given an open addressing hash table T with a fixed load factor of $\alpha : 1 > \alpha > 0$, assume T is filled using double hashing to load factor α. As discussed in subsection 2.3, double hashing uses two hash functions h_1 and h_2. The goal of this section is to show that if both hash functions h_1 and h_2 are randomly, uniformly, and independently chosen from the strongly universal hash functions H (Definition 6), then the expected cost of an unsuccessful search using double hashing is $\frac{1}{1-\alpha}$ table accesses. This question was suggested by Carter and Wegman [8].

Another important form of hashing is *hashing with chaining*; see, for example, [11, 13, 24]. Carter and Wegman [8] showed that given any strongly universal set of hash functions H, then randomly and uniformly selecting a hash function $h \in H$ gives an expected chain length of at most $1 + \alpha'$ for fixed load factor $\alpha' > 0$. For instance, taking the set of hash functions H with domain and range $\mathbb{Z}_p^+$ as in Definition 6, Carter and Wegman's result is important since the strongly universal set H (of size $O(p^2)$) behaves as if randomly selecting a function from the set of all functions from $\mathbb{Z}_p^+$ to $\mathbb{Z}_p^+$ (of size $O(p^p)$). See Mehlhorn [24, 23] for lower bounds on the sizes of universal hash sets.

As future research they suggest extending such an analysis to double or open hashing. Schmidt and Siegel [30] and Siegel [31] answer this, giving $c \log n$-independent functions that are computable in constant time for a standard word model random access machine. Their results are quite general; see also [32]. Next, we focus on another answer to Carter and Wegman's question using the standard set H of strongly universal hash functions, see Definition 6, as they are represented in the G_3 graph. Although this paper uses a different model, the $\boldsymbol{g}$-edges in G_3 make selecting entire blocks simulate twowise independent functions; see Theorem 3.

The G_3 graph can represent a double hashing configuration if all elements in each input block are either all selected or all unselected. That is, say each block $|\widehat{I}_j|/p = \alpha_j$ is such that either $\alpha_j = 1$ or $\alpha_j = 0$. This gives a fixed load factor $\alpha = |\widehat{\mathcal{I}}|/|\mathcal{I}|$, where $1 > \alpha > 0$. Each entire input block corresponds to a cell in the hash table T, where T is of size p. So, if $\alpha_j = 1$, then $T[j]$ is full; and if $\alpha_j = 0$, then $T[j]$ is empty.

If one wants to build a double hashing table, do this by making two independent and uniform random choices $h_1, h_2 \in H$, where H is the strongly universal set described in subsection 2.4. So, given a key x, the value $j_0 = h_1(x)$ is the first table element $T[j_0]$ to probe. Now, if $T[j_0]$ is full and $T[j_0] \neq x$, then probe the values $T[(j_0 + i\, h_2(x)) \bmod p]$ for $i = 1, \ldots, p-1$, until encountering the key x or an empty table element (NIL).

Next, this paper shows that building a hash table by double hashing with the initial uniform and independent random choices $h_1, h_2 \in H$ gives an open addressing table of load factor α with expected number of probes $\frac{1}{1-\alpha}$ for an unsuccessful search, regardless of the input distribution.

When searching through a hash table for x, a collision sequence equates to a probe sequence $(j_0 + i\, h_2(x)) \bmod p$, given $j_0 \leftarrow h_1(x)$ and h_1, h_2 both randomly, uniformly, and independently chosen from a strongly universal set H.

In this context, consider the collision sequence $\mathcal{C}_{J_1,K}$ starting in position (J_1, K). So let J_1 and K be randomly, uniformly, and independently chosen. Since J_1 is independent of J_2 and $(J_2, K) = \boldsymbol{g}(J_1, K)$, then

$$\begin{aligned}&\mathbf{Pr}[\mathbf{Length}(\mathcal{C}_{J_1,K}) \geq 2]\\ &= \mathbf{Pr}[\mathbf{Length}(\mathcal{C}_{J_1,K}) = 1 \wedge \mathbf{Length}(\mathcal{C}_{J_2,K}) \geq 1 \wedge (J_2, K) = \boldsymbol{g}(J_1, K)].\end{aligned}$$

In the next proof, lower-order terms that would appear if the adversary selected

degenerate elements are ignored.

PROPOSITION 1 (hashing collision sequence). *Say T is an open address hash table with any configuration of elements built by initially randomly, uniformly, and independently choosing h_1, h_2 from H and then performing double hashing to insert elements into T. Let $\alpha = |\widehat{\mathcal{I}}|/|\mathcal{I}|$. Assume each block I_j is such that $|\widehat{I}_j|/p = \alpha_j$ and either $\alpha_j = 1$ or $\alpha_j = 0$. Then the expected collision sequence length in T is* $\mathbf{E}[\mathbf{Length}(\mathcal{C})] \leq \frac{\alpha}{1-\alpha}$.

Proof. For any fixed $\mathcal{S} = \{\alpha_{i_0}, \ldots, \alpha_{i_s}\} \subset \{\alpha_0, \ldots, \alpha_{p-2}\}$ let $s < p-2$ and $\alpha_{i_r} = 1$ for all $i_r \in \{i_0, \ldots, i_s\}$ and $\alpha_j = 0$ for all $j \notin \{i_0, \ldots, i_s\}$. This means $\alpha = \frac{s+1}{p}$. In this case, since each full block ($\alpha_{i_r} = 1$) has exactly one edge going to all other blocks, then the fraction $1 - \frac{s+1}{p}$ of all collision sequences starting in any full blocks are of length exactly 1.

Now the claim that $\mathbf{E}[\mathbf{Length}(\mathcal{C})] \leq \frac{\alpha}{1-\alpha}$ is shown by induction.

Let $\mathcal{C}_{J_1,K}$ be a potential collision sequence that passes through at any randomly and uniformly chosen $(J_1, K) \in \mathbb{Z}_p^+ \times \mathbb{Z}_p^+$; then $\mathbf{E}[\mathbf{Pr}[\mathbf{Length}(\mathcal{C}_{J_1,K}) \geq 1]] = \alpha$ since α is the probability of randomly and uniformly selecting an input element.

Basis. Since $\alpha = \frac{s+1}{p}$, then randomly and uniformly choosing $(J_1, K) \in \mathbb{Z}_p^+ \times \mathbb{Z}_p^+$, where $(J_2, K) = \boldsymbol{g}(J_1, K)$, gives

$$\begin{aligned}\mathbf{Pr}[(J_1, K) \in \widehat{I}_{J_1} \wedge (J_2, K) \in \widehat{I}_{J_2}] &= \left(\frac{s+1}{p}\right)\left(\frac{s}{p}\right) \\ &\leq \alpha^2\end{aligned}$$

by Lemma 2 noting that since $\alpha_{i_r} = 1$, for all $\alpha_{i_r} \in \mathcal{S}$, then $n_0 = \cdots = n_{p-1} = \alpha\, p$. Further, if $T[J_1]$ is full, then $\mathbf{Pr}[(J_2, K) \in \widehat{I}_{J_2}] = \alpha - \frac{1}{p} = \frac{s}{p}$ and $\frac{s}{p} < \alpha$.

Inductive hypothesis. For some $c \geq 2$, and for all $i < c$, assume for any collision sequences $\mathcal{C}_{J_1,K}$ that $\mathbf{Pr}[\mathbf{Length}(\mathcal{C}_{J_1,K}) \geq i] \leq \alpha^i$.

Inductive step. Take $c \geq 2$ and consider $t \leq c$; then we claim for all potential collision sequences $\mathcal{C}_{J_1,K}$, where $(J_2, K) = \boldsymbol{g}(J_1, K)$,

$$\mathbf{Pr}[\mathbf{Length}(\mathcal{C}_{J_1,K}) \geq t+1] = \mathbf{Pr}[\mathbf{Length}(\mathcal{C}_{J_1,K}) = 1]\ \mathbf{Pr}[\mathbf{Length}(\mathcal{C}_{J_2,K}) \geq t].$$

To substantiate this claim, take a *potential* length $t+1$ collision sequence $\mathcal{C}_{J_1,K}$ and suppose the first probe starts in block I_{J_1}, and so

$$\mathcal{C}_{J_1,K} = (J_1, K) \to (J_2, K) \to \cdots \to (J_t, K)$$

and $J_i = \boldsymbol{g}_1(J_{i-1}, K)$, where $c \geq t \geq i > 1$.

It must be that

$$\begin{aligned}\mathbf{Pr}[\mathbf{Length}(\mathcal{C}_{J_1,K}) \geq t+1] &= \mathbf{Pr}[\mathbf{Length}(\mathcal{C}_{J_1,K}) = 1]\ \mathbf{Pr}[\mathbf{Length}(\mathcal{C}_{J_2,K}) \geq t] \\ &\leq \alpha\ \mathbf{Pr}[\mathbf{Length}(\mathcal{C}_{J_2,K}) \geq t],\end{aligned}$$

which holds by pairwise independence from Theorem 3 and, further, by the inductive hypothesis $\mathbf{Pr}[\mathbf{Length}(\mathcal{C}_{J_2,K}) \geq t] \leq \alpha^t + \epsilon$, completing the induction.

Now, for any $t \geq 1$, since the random variable $\mathbf{Length}(\mathcal{C}_{J_1,K})$ is nonnegative, this means

$$\begin{aligned}\mathbf{E}[\mathbf{Length}(\mathcal{C})] &= \sum_{t \geq 1} \mathbf{Pr}[\mathbf{Length}(\mathcal{C}) \geq t] \\ &\leq \sum_{t \geq 1} \alpha^t \\ &\leq \frac{\alpha}{1-\alpha}.\end{aligned}$$

This completes the proof. □

Suppose $t < p-1$ elements are in an open addressing hashing table T, where $|T| = p$, that is filled with $t = \alpha\, p$ elements. Such a configuration represents the elements of T that are filled with load factor $\alpha = \frac{t}{p}$. In addition, by Proposition 1, the expected collision sequence length is $\frac{\alpha}{1-\alpha}$, no matter how the table T is filled. In the next theorem, α is the load factor of the table T.

THEOREM 8 (main double hashing theorem). *Say T is an open address hash table with any configuration of elements built by initially randomly, uniformly, and independently choosing h_1, h_2 from H and then performing double hashing to insert elements into T. Now an unsuccessful search for a key x in T using double hashing with h_1 and h_2 has expected cost of at most $1 + \mathbf{E}[\mathbf{Length}(\mathcal{C})] = \frac{1}{1-\alpha}$ hash probes.*

Proof. Suppose $t = \alpha\, p$ different keys $x_1, \ldots, x_t$ have been inserted into the table T using the randomly, independently, and uniformly chosen $h_1, h_2 \in H$. Then there are $t = \alpha\, p$ blocks $I_{j_1}, \ldots, I_{j_t}$ that have $\alpha_{j_r} = 1$ for all $j_r \in \{j_1, \ldots, j_t\}$. That is, the table T has load factor α. Note that $\alpha_{j'_r} = 0$ for all $j'_r \in \mathbb{Z}_p^+ - \{j_1, \ldots, j_t\}$.

Now suppose we are searching for a key $x \notin \{x_1, \ldots, x_t\}$ in T given the hash functions h_1, h_2. Since H is strongly universal, it must be that for any $x_i \in \{x_1, \ldots, x_t\}$, then

$$\mathbf{Pr}[h_1(x) = h_1(x_i)] = \frac{1}{p},$$
$$\mathbf{Pr}[h_2(x) = h_2(x_i)] = \frac{1}{p}.$$

This means, for $x_i \in \{x_1, \ldots, x_t\}$ and $x \notin \{x_1, \ldots, x_t\}$, that

$$\begin{aligned}\mathbf{Pr}[h_1(x) = h_1(x_i) \wedge h_2(x) = h_2(x_i)] &= \mathbf{Pr}[h_1(x) = h_1(x_i)]\ \mathbf{Pr}[h_2(x) = h_2(x_i)] \\ &= \frac{1}{p(p-1)}.\end{aligned}$$

Therefore, the probing sequence for x has equal probability $(\frac{1}{p(p-1)})$ of starting at any $(J_1, K) \in \mathbb{Z}_p^+ \times (\mathbb{Z}_{p-1}^+ - \{0\})$. Thus, applying Proposition 1, the expected collision sequence length is $\frac{\alpha}{1-\alpha} + 1 = \frac{1}{1-\alpha}$ for any $x \notin \{x_1, \ldots, x_t\}$ and x not in T.

This completes the proof. □

This last theorem assumes any (j, k) is not of the form $(j, 0)$ for any $j \in \mathbb{Z}_p^+$. Allowing such elements to be selected from $\mathbb{Z}_p^+ \times \mathbb{Z}_{p-1}^+$ would add a lower-order term of $O(\frac{1}{p})$ to $\frac{1}{1-\alpha}$.

4. Showing the graph G_3 expands. Suppose $\mathcal{L}(\widehat{I}_j) \leq 1$ for all $j \in \mathbb{Z}_p^+$. By Theorem 6 for randomly and uniformly chosen $(J_1, K) \in \mathbb{Z}_p^+ \times \mathbb{Z}_p^+$, then

$$\mathbf{Pr}[(J_1, K) \in \widehat{I}_{J_1} \wedge (J_2, K) \in \widehat{I}_{J_2}] \leq \alpha^2 + \epsilon,$$

where $(J_2, K) = \boldsymbol{g}(J_1, K)$ and ϵ is a lower-order term.

Let L_1 be the set of collision sequences of length at least 1. Suppose $(J_1, K) \in \mathbb{Z}_p^+ \times \mathbb{Z}_p^+$ is randomly and uniformly chosen and it so happens that $(J_1, K) \in L_1$; then $\mathbf{Pr}[\mathbf{Length}(\mathcal{C}_{J_1,K}) \geq 1 \mid (J_1, K) \in L_1] = 1$, and $\mathbf{Pr}[\mathbf{Length}(\mathcal{C}_{J_1,K}) \geq 2 \mid (J_1, K) \in L_1] = \alpha$. This last equality holds because $(J_1, K) = \boldsymbol{g}(J_0, K)$ and

$$\begin{aligned}\mathbf{Pr}[(J_1, K) \in \widehat{I}_{J_1} \wedge (J_2, K) \in \widehat{I}_{J_2} \mid (J_1, K) \in \widehat{I}_{J_1}] &= \mathbf{Pr}[(J_2, K) \in \widehat{I}_{J_2} \mid (J_1, K) \in \widehat{I}_{J_1}] \\ &= \mathbf{Pr}[(J_2, K) \in \widehat{I}_{J_2}],\end{aligned}$$

and the last equality is by pairwise independence of Theorem 3. Furthermore, $\mathbf{Pr}[(J_2, K) \in \widehat{I}_{J_2}] = \alpha$ since (J_2, K) is independent of (J_1, K), making (J_2, K) randomly and uniformly chosen from $\mathbb{Z}_p^+ \times \mathbb{Z}_p^+$.

In the next proof, lower-order terms that would appear if the adversary selected degenerate elements are ignored.

PROPOSITION 2 (general collision sequence length). *Take G_3 so that $\alpha = |\widehat{\mathcal{I}}|/|\mathcal{I}|$, where α is fixed and $1 > \alpha > 0$, and $\mathcal{L}(\widehat{I}_j) \leq 1$ for all $j \in \mathbb{Z}_p^+$; then the expected collision sequence length is $\mathbf{E}[\mathbf{Length}(\mathcal{C})] \leq 1 + \frac{2\,\alpha}{1-\alpha}$.*

Proof. The fact that a collision sequence traveling through $(J_1, K) \in \widehat{I}_{J_1}$ has expected remaining length via $\boldsymbol{g}$ of $\mathbf{E}[\mathbf{Length}(\mathcal{C}^r_{J_1,K})] \leq \frac{\alpha}{1-\alpha}$ is proved by induction. After the induction, the proof accounts for the expected prior collision sequence length $\mathbf{E}[\mathbf{Length}(\mathcal{C}^p_{J_1,K})]$ going to (J_1, K). (Note that $\mathcal{C}^p$ is *prior* collision sequence and $\mathcal{C}^r$ is the *remaining* collision sequence.)

Basis. Assuming $(J_1, K) \in \mathbb{Z}_p^+ \times \mathbb{Z}_p^+$ is randomly and uniformly chosen and $(J_1, K) \in \widehat{I}_{J_1}$ and $(J_2, K) = \boldsymbol{g}(J_1, K)$, then

$$\begin{aligned}
&\mathbf{Pr}[\mathbf{Length}(\mathcal{C}^r_{J_1,K}) \geq 2 \mid (J_1, K) \in \widehat{I}_{J_1}] \\
&= \mathbf{Pr}[\mathbf{Length}(\mathcal{C}^r_{J_1,K}) \geq 1 \wedge \mathbf{Length}(\mathcal{C}^r_{J_2,K}) \geq 1 \mid (J_1, K) \in \widehat{I}_{J_1}] \\
&= \mathbf{Pr}[(J_2, K) \in \widehat{I}_{J_2}] \\
&\leq \alpha,
\end{aligned}$$

by the pairwise independence of J_1 and J_2 and the uniformity of (J_2, K), since $(J_2, K) = \boldsymbol{g}(J_1, K)$ by Theorem 3 and by the choice of (J_1, K).

Inductive hypothesis. For some $c \geq 2$, and for all $i < c$, assume for any potential collision sequence $\mathcal{C}^r_{J_1,K}$ starting in block I_{J_1}, so that $(J_1, K) \in \widehat{I}_{J_1}$, and traveling via $\boldsymbol{g}$ that $\mathbf{Pr}[\mathbf{Length}(\mathcal{C}^r_{J_1,K}) \geq i \mid (J_1, K) \in \widehat{I}_{J_1}] \leq \alpha^{i-1} + \epsilon$ for $i > 1$ and for $i = 1$; then $\epsilon = 0$.

Inductive step. Take $c \geq 2$ and consider $t \leq c$; then take a potential length $t + 1$ collision sequence starting with $(J_1, K) \in \widehat{I}_{J_1}$ and traveling via $\boldsymbol{g}$ so for $i : t \geq i \geq 0$,

$$\mathcal{C}^r_{J_1,K} = (J_1, K) \to (J_2, K) \to \cdots \to (J_t, K)$$

and $J_i = \boldsymbol{g}_1(J_{i-1}, K)$, where $c \geq t \geq i > 1$ and $(J_1, K) \in \mathbb{Z}_p^+ \times \mathbb{Z}_p^+$. It must be that

$$\begin{aligned}
&\mathbf{Pr}[\mathbf{Length}(\mathcal{C}^r_{J_1,K}) \geq t + 1 \mid (J_1, K) \in \widehat{I}_{J_1}] \\
&= \mathbf{Pr}[\mathbf{Length}(\mathcal{C}^r_{J_1,K}) \geq 1]\ \mathbf{Pr}[\mathbf{Length}(\mathcal{C}^r_{J_2,K}) \geq t] \\
&= \mathbf{Pr}[\mathbf{Length}(\mathcal{C}^r_{J_2,K}) \geq t],
\end{aligned}$$

which holds by pairwise independence by Theorem 3 and by Theorem 6 and, further, by the inductive hypothesis $\mathbf{Pr}[\mathbf{Length}(\mathcal{C}^r_{J_2,K}) \geq t] \leq \alpha^{t-1} + \epsilon$, completing the induction.

Now, without conditioning on the first element of a collision sequence being selected, say for any $t \geq 1$, since the random variable $\mathbf{Length}(\mathcal{C}^r_{J_1,K})$ is nonnegative,

this means

$$\begin{aligned}\mathbf{E}[\mathbf{Length}(\mathcal{C}^r)] &= \sum_{t\geq 1} \mathbf{Pr}[\mathbf{Length}(\mathcal{C}^r) \geq t] \\ &\leq \sum_{t\geq 1} \alpha^t \\ &\leq \frac{\alpha}{1-\alpha}.\end{aligned}$$

Therefore, conditioning on the first element of a collision sequence being selected gives a bound of $1 + \frac{\alpha}{1-\alpha}$.

The induction above is based on starting at a random uniformly chosen (J_1, K) and going via $\boldsymbol{g}$. The issue of the prior expected collision sequence length $\mathbf{E}[\mathbf{Length}(\mathcal{C}^p_{J_1,K})] = \frac{\alpha}{1-\alpha}$ is dealt with by a symmetric argument. This means the expected collision sequence length is

$$\begin{aligned}\mathbf{E}[\mathbf{Length}(\mathcal{C}_{J_1,K})] &= \mathbf{E}[\mathbf{Length}(\mathcal{C}^p_{J_1,K})] + \mathbf{E}[\mathbf{Length}(\mathcal{C}^r_{J_1,K})] \\ &= \frac{2\,\alpha}{1-\alpha}.\end{aligned}$$

Finally, conditioning on the first element of a collision sequence being selected gives $1 + \frac{2\,\alpha}{1-\alpha}$.

This completes the proof. □

Proposition 2 says for any G_3 where $\mathcal{L}(\widehat{I}_j) \leq 1$, for all $j \in \mathbb{Z}_p^+$, then the expected length of collision sequences is $1 + \frac{2\alpha}{1-\alpha}$. Thus randomly and uniformly selecting a collision sequence $\mathcal{C}$ from such a G_3, then this collision sequence will be of expected length $1 + \frac{2\alpha}{1-\alpha}$. This means all collision sequences' lengths added together divided by the number of collision sequences is bounded above by $1 + \frac{2\alpha}{1-\alpha}$.

LEMMA 7. *Suppose $\mathcal{L}(\widehat{I}_j) \leq 1$, for all $j \in \mathbb{Z}_p^+$, and $\alpha = |\widehat{\mathcal{I}}|/|\mathcal{I}|$. Then*

$$\mathcal{G}(\widehat{\mathcal{I}}) \geq \frac{1-\alpha}{1+\alpha}|\widehat{\mathcal{I}}|.$$

Proof. Let c be the total number of collision sequences of length greater than 0. Also, the ith collision sequence is Coll_Sequence(i) for $i : c \geq i \geq 1$. A randomly and uniformly selected collision sequence from G_3 when $\mathcal{L}(\widehat{I}_j) \leq 1$ for $j \in \mathbb{Z}_p^+$ has expected length $1 + \frac{2\,\alpha}{1-\alpha}$ by Proposition 2. This means the average of all the collision sequence lengths is $1 + \frac{2\,\alpha}{1-\alpha}$. Thus, if $\mathcal{C}$ represents a randomly and uniformly chosen collision sequence, then

$$\mathbf{E}[\mathbf{Length}(\mathcal{C})] = \frac{\sum_{i=1}^{c} \mathbf{Length}(\text{Coll\_Sequence}(i))}{c}.$$

Furthermore, since $\mathbf{E}[\mathbf{Length}(\mathcal{C})] \leq 1 + \frac{2\,\alpha}{1-\alpha}$ by Proposition 2 and since $\mathcal{L}(\widehat{I}_j) \leq 1$, for all $j \in \mathbb{Z}_p^+$, and since

$$|\widehat{\mathcal{I}}| = \sum_{i=1}^{c} \mathbf{Length}(\text{Coll\_Sequence}(i)),$$

it must be that

$$\frac{|\widehat{\mathcal{I}}|}{c} \leq 1 + \frac{2\alpha}{1-\alpha}.$$

Finally, each of the collision sequences has global expansion of 1, and so $\mathcal{G}(\widehat{\mathcal{I}}) = c$. In other words,

$$\frac{1-\alpha}{1+\alpha}|\widehat{\mathcal{I}}| \leq \mathcal{G}(\widehat{\mathcal{I}}).$$

This completes the proof. □

The function $\frac{1-\alpha}{1+\alpha}$ minimizes at $\alpha = \frac{1}{2}$, for $\alpha : \frac{1}{2} \geq \alpha > 0$, since

$$\frac{d}{d\alpha}\left(\frac{1-\alpha}{1+\alpha}\right) = \frac{-1}{1+\alpha} - \frac{1-\alpha}{(1+\alpha)^2},$$

which is negative for $\alpha : \frac{1}{2} \geq \alpha > 0$.

DEFINITION 13. *Two collision sequences*

$$\mathcal{C}_1 = (j_1, k) \to (j_2, k) \to \cdots \to (j_t, k),$$
$$\mathcal{C}_2 = (j_{t+2}, k) \to (j_{t+3}, k) \to \cdots \to (j_u, k)$$

can be joined *into one by selecting the element* (j_{t+1}, k)*, where*

$$g(j_t, k) \to (j_{t+1}, k) \quad \textit{and}$$
$$g(j_{t+1}, k) \to (j_{t+2}, k).$$

In this case, $\mathcal{C}_1$ *and* $\mathcal{C}_2$ *are* separated *by one selection.*

Consider a total of $c < p/2$ collision sequences,

$$\mathcal{C}_1,\ \mathcal{C}_2, \ldots, \mathcal{C}_c,$$

where $\mathcal{C}_i$ and $\mathcal{C}_{i+1}$, for $i : c > i \geq 1$, are separated by one selection. Then selecting $c-1$ elements joins all of these collision sequences into one single collision sequence. This can be stated as the following lemma.

LEMMA 8. *Given* $c < p/2$ *collision sequences* $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_c$*, where* $\mathcal{C}_i$ *and* $\mathcal{C}_{i+1}$*, for* $i : c > i \geq 1$*, are each separated by one selection, then* $c-1$ *singletons can be used to join these collision sequences into one single collision sequence.*

THEOREM 9 (main expander theorem). *Consider any input set* $\widehat{\mathcal{I}} \subset \mathcal{I}$ *from* G_3*, where* $|\widehat{\mathcal{I}}| \leq \frac{p^2}{2}$ *so* $\alpha \leq \frac{1}{2}$*. The subgraph* G_3 *expands by at least* $\frac{3}{16}$*.*

Proof. Suppose the adversary chooses the elements $\widehat{\mathcal{I}}$ from $\mathcal{I}$, where $|\widehat{\mathcal{I}}| \leq \frac{1}{2}|\mathcal{I}|$.

If $\mathcal{L}(\widehat{\mathcal{I}}) \geq \frac{3}{16}|\widehat{\mathcal{I}}|$, then we are done since there is a total of at least $\frac{3}{16}$ expansion. Therefore, consider the case where $\mathcal{L}(\widehat{\mathcal{I}}) < \frac{3}{16}|\widehat{\mathcal{I}}|$.

Start with the situation where $\mathcal{L}(\widehat{I}_j) \leq 1$ for all $j \in \mathbb{Z}_p^+$; then by Lemma 7 there is global expansion of at least $\frac{1}{3}$ since $\alpha = \frac{1}{2}$ minimizes $\frac{1-\alpha}{1+\alpha}$ for $\alpha : \frac{1}{2} \geq \alpha > 0$. If $\alpha = \frac{1}{2}$, then $\frac{1-\alpha}{1+\alpha} = \frac{1}{3}$, and so

$$\mathcal{G}(\widehat{\mathcal{I}}) \geq \frac{1}{3}|\widehat{\mathcal{I}}|.$$

Now take the more general situation where $\widehat{\mathcal{I}}$ is such that $\mathcal{L}(\widehat{\mathcal{I}}) < \frac{3}{16}|\widehat{\mathcal{I}}|$. There are several cases to consider.

- *Case* 1. Not counting singletons.
  For now, assume more than $\frac{13}{16}|\widehat{\mathcal{I}}|$ selected elements are in ***loc***-contiguous subblocks. This case assumes no singletons. Thus, ignore the up to $\frac{3}{16}|\widehat{\mathcal{I}}|$ potential singletons.
  Therefore, applying Lemma 7 with $\alpha' \leq \frac{13}{16}\alpha \leq \frac{13}{32}$, since $\alpha \leq \frac{1}{2}$, gives

$$\begin{aligned}\mathcal{G}(\widehat{\mathcal{I}}) &\geq \frac{1-\frac{13}{32}}{1+\frac{13}{32}}|\widehat{\mathcal{I}}| \\ &= \frac{19}{45}|\widehat{\mathcal{I}}|.\end{aligned}$$

  Furthermore, the function $\frac{1-\alpha'}{1+\alpha'}$ is minimal for $\alpha' = \frac{13}{32}$, where $\alpha' : \frac{13}{32} \geq \alpha' > 0$. This case is complete since $\frac{19}{45} > \frac{3}{16}$.
- *Case* 2. Extending collision sequences.
  First, if the up to $\frac{3}{16}|\widehat{\mathcal{I}}|$ of the elements are used to extend but not join any two collision sequences, then the global expansion remains the same. Here the adversary has simply made the collision sequences longer, thereby changing in which block each of them terminates. But the same number of collision sequence ends remain, giving the same global expansion. This case is complete.
- *Case* 3. Joining collision sequences.
  Suppose each of the up to $\frac{3}{16}|\widehat{\mathcal{I}}|$ singleton elements can be used to join collision sequences together.
  Of course, it is given there are at least

$$\frac{19}{45}|\widehat{\mathcal{I}}| > \frac{6}{16}|\widehat{\mathcal{I}}|$$

  collision sequences by Case 1 (by applying Lemma 7). If each of the $\frac{3}{16}|\widehat{\mathcal{I}}|$ singleton elements joins exactly two collision sequences and all such pairs of joined collision sequences are disjoint, then the global expansion is reduced by at most $\frac{3}{16}$. This is because joining two collision sequences reduces the global expansion by one. Now if two singletons join three collision sequences, then the global expansion is also reduced by two. Lemma 8 indicates that each singleton reduces the global expansion by exactly one.
  Taking this argument to its logical end, let $\frac{3}{16}|\widehat{\mathcal{I}}|$ singleton elements be used to join at most $\frac{3}{16}|\widehat{\mathcal{I}}| + 1$ collision sequences. Applying Lemma 8, the $\frac{3}{16}|\widehat{\mathcal{I}}|$ singletons can be used to join as many as $\frac{3}{16}|\widehat{\mathcal{I}}| + 1$ collision sequences into one or a few collision sequences. Suppose these collision sequences do not have any expansion (i.e., they are of length p). Going further, say adding these singletons forms new ***loc***-contiguous subblocks and does not add any expansion themselves.
  This leaves more than

$$\frac{6}{16}|\widehat{\mathcal{I}}| - \frac{3}{16}|\widehat{\mathcal{I}}| = \frac{3}{16}|\widehat{\mathcal{I}}|$$

  collision sequences, giving global expansion of at least $\frac{3}{16}$, completing this case.

This completes the proof. □

Next, an accounting is made for an adversary selecting degenerate elements $(j,0)$ for all $j \in \mathbb{Z}_p^+$ and $(p-1,k)$ for all $k \in \mathbb{Z}_p^+$. Suppose the elements $(j,0)$ for all $j \in \mathbb{Z}_p^+$ are selected. Recall these degenerate elements give no global expansion since $\boldsymbol{g}(j,0) = \boldsymbol{id}(j,0)$. Further, assume they extend ***loc***-contiguous sequences in each $\widehat{I}_j$, thus giving no additional local expansion. Thus, in this case the selection of these elements takes the expansion from $\frac{3}{16}$ to $\frac{3}{16}(\frac{1}{1+\frac{1}{p}})$.

In addition, suppose the elements $(p-1,k)$ for all $k \in \mathbb{Z}_p^+$ are also selected. These elements give no local expansion since $\boldsymbol{loc}(p-1,k) = (p-1,k)$. However, it is possible that $p-1$ of these elements can join two collision sequences together. Note that node $(p-1,0)$ is degenerate both locally and globally. Joining pairs of collision sequences together without adding local expansion changes the expansion of $\frac{3}{16}(\frac{1}{1+\frac{1}{p}})$ to

$$\frac{3}{16}\left(\frac{1}{1+\frac{1}{p}}\right)\left(1-\frac{1}{p-1}\right) = \frac{3}{16}\left(\frac{p}{p+1}\right)\left(\frac{p-2}{p-1}\right).$$

5. Conclusions. This paper gives a new way of showing expansion of three permutations comprising the Gabber–Galil expander. This is done without using eigenvalue methods or higher algebra. We use methods based on Chor and Goldreich's Theorem 3 on pairwise independence. It is important to notice that we are applying the probabilistic method to a fixed graph. For $\alpha = \frac{1}{2}$, Theorems 8 and 9 tie closely the expected collision sequence length of $2 = \frac{2\alpha}{1-\alpha} = \frac{1}{1-\alpha}$, giving insight into double hashing and graph expansion. This is interesting since our expander result says no matter what distribution of inputs are chosen while restricting local expansion, then we still have expected collision sequence length bounded by 2. If local expansion is not restricted sufficiently, then the graph expands (locally) as well. Likewise, for double hashing, no matter what input distribution is assumed for the keys, randomly, independently, and uniformly choosing two universal hash functions gives expected collision sequence length bounded by 2. This gives insight into both graph expanders as well as double hashing.

Fundamentally, universal hash functions are small sets that "randomize" well. Likewise, expander graphs have relatively few edges, yet they seem to have many properties amenable to randomness.

Acknowledgment. We are grateful to a referee for pointing out a problem in a previous version.

REFERENCES

[1] M. Ajtai, *Recursive construction for* 3*-regular expanders*, Combinatorica, 14 (1994), pp. 379–416.
[2] N. Alon, *Eigenvalues and expanders*, Combinatorica, 6 (1986), pp. 83–96.
[3] N. Alon, *Tools from higher algebra*, in Handbook of Combinatorics, Vol. 1, 2, R. L. Graham, M. Grötschel, and L. Lovász, eds., Elsevier, Amsterdam, 1995, pp. 1749–1783.
[4] N. Alon, Z. Galil, and V. D. Milman, *Better expanders and superconcentrators*, J. Algorithms, 8 (1987), pp. 337–347.
[5] N. Alon and J. Spencer, *The Probabilistic Method*, Wiley-Interscience, John Wiley and Sons, New York, 1992.
[6] M. Blum, R. M. Karp, O. Vornberger, C. H. Papadimitriou, and M. Yannakakis, *The complexity of testing whether a graph is a superconcentrator*, Inform. Process. Lett., 13 (1981) pp. 164–167.
[7] M. Capalbo, O. Reingold, S. Vadhan, and A. Wigderson, *Randomness conductors and constant-degree lossless expanders*, in Proceedings of the Thirty-Fourth Annual Symposium on Theory of Computing, ACM, New York, 2002, pp. 659–668.

[8] J. L. Carter and M. N. Wegman, *Universal classes of hash functions*, J. Comput. System Sci., 18 (1979), pp. 143–154.
[9] B. Chor and O. Goldreich, *On the power of two-points based sampling*, J. Complexity, 5 (1989), pp. 96–106.
[10] F. R. K. Chung, *Spectral Graph Theory*, CBMS Regional Conf. Ser. in Math. 92, AMS, Providence, RI, 1997.
[11] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd ed., MIT Press, Cambridge, MA, 2001.
[12] O. Gabber and Z. Galil, *Explicit construction of linear-sized superconcentrators*, J. Comput. System Sci., 22 (1981), pp. 407–420.
[13] G. H. Gonnet and R. A. Baeza-Yates, *Handbook of Algorithms and Data Structures*, 2nd ed., Addison–Wesley, Reading, MA, 1990.
[14] L. Guibas and E. Szemerédi, *The analysis of double hashing*, J. Comput. System Sci., 16 (1978), pp. 226–274.
[15] S. Jimbo and A. Maruoka, *Expanders obtained from affine transformations*, Combinatorica, 7 (1987), pp. 343–355.
[16] N. Kahale, *Eigenvalues and expansion of regular graphs*, J. Assoc. Comput. Mach., 42 (1995), pp. 1091–1106.
[17] D. E. Knuth, *The Art of Computer Programming*, *Sorting and Searching*, Vol. 3, Addison–Wesley, Reading, MA, 1973.
[18] A. Lubotzky, *Discrete Groups, Expanding Graphs, and Invariant Measures*, Progr. Math. 125, Birkhäuser, Basel, 1994.
[19] A. Lubotzky, R. Phillips, and P. Sarnak, *Ramanujan graphs*, Combinatorica, 8 (1988), pp. 261–277.
[20] G. S. Lueker and M. Molodowitch, *More analysis of double hashing*, Combinatorica, 13 (1993), pp. 83–96.
[21] G. A. Margulis, *Explicit constructions of concentrators*, Problemy Peredachi Informacii, 9 (1973), pp. 71–80 (in Russian); Problems Inform. Transmission, 9 (1975), pp. 325–332 (in English).
[22] G. A. Margulis, *Explicit group-theoretical constructions of combinatorial schemes and their application to the design of expanders and superconcentrators*, Problemy Peredachi Informatsii, 24 (1988), pp. 51–60 (in Russian); Problems Inform. Transmission, 24 (1988), pp. 39–46 (in English).
[23] K. Mehlhorn, *On the program size of perfect and universal hash functions*, in Proceedings of the 23rd Annual Symposium on Foundations of Computer Science, IEEE, New York, 1982, pp. 170–175.
[24] K. Mehlhorn, *Data Structures and Efficient Algorithms*, Springer-Verlag, Berlin, 1984.
[25] R. Meshulam and A. Wigderson, *Expanders from symmetric codes*, in Proceedings of the Thirty-Fourth Annual Symposium on Theory of Computing, ACM, New York, 2002, pp. 669–677.
[26] R. Motwani and P. Raghavan, *Randomized Algorithms*, Cambridge University Press, Cambridge, UK, 1995.
[27] M. Pinsker, *On the complexity of a concentrator*, in Proceedings of the 7th International Teletraffic Conference, Stockholm, Sweden, 1973.
[28] O. Reingold, S. Vadhan, and A. Wigderson, *Entropy waves, the zig-zag graph product, and new constant-degree expanders and extractors*, Ann. of Math. (2), 155 (2002), pp. 157–187.
[29] P. Sarnak, *Some Applications of Modular Forms*, Cambridge Tracts in Math. 99, Cambridge University Press, Cambridge, UK, 1990.
[30] J. P. Schmidt and A. Siegel, *Double Hashing Is Computable and Randomizable with Universal Hash Functions*, NYU Technical report TR1995-686, New York University, New York, NY, 1995.
[31] A. Siegel, *On universal classes of fast hash functions, their time-space tradeoff, and their applications*, in Proceedings of 30th Annual Symposium on Foundations of Computer Science, IEEE, New York, 1989, pp. 20–25.
[32] A. Siegel, *On universal classes of extremely random constant-time hash functions*, SIAM J. Comput., 33 (2004), pp. 505–543.
[33] A. Wigderson and D. Zuckerman, *Expanders that beat the eigenvalue bound: Explicit construction and application*, Combinatorica, 19 (1999), pp. 125–138.