Identity-Free Set Systems

Research Capstone Project

Shyamal Patel
Advised by Dr. Greg Bodwin

Faculty Mentor: Professor Santosh Vempala

Second Reader: Professor Ernest Croot

College of Computing
Georgia Tech
Atlanta, GA
April 22nd, 2020

Acknowledgements

First and foremost, I would like to thank my advisor, Dr. Greg Bodwin, without whom this thesis would not exist. He's given me countless pieces of advice and has introduced me to several beautiful problems about shortest paths and extremal combinatorics. I would also like to thank Professor Santosh Vempala and Professor Ernest Croot for agreeing to read to my thesis.

Next, I would like to thank Professor Aaron Bernstein, Dr. Fidel Barrera-Cruz, Professor Grigoriy Blekherman, and Samantha Petti for working with me over the last four years and teaching me about research. I am particularly grateful to Dr. Fidel Barrera-Cruz and Samantha Petti for starting me on the path of research and introducing me to extremal and probabilistic combinatorics and theoretical computer science.

I would also like to extend thanks to all my friends at Georgia Tech, who made the last four years so enjoyable. I'll definitely always remember my late-night sublime runs and wall scaling with Daniel Hathcock and Sherry Sarkar as well and as the games of charades and duck tape ring toss with Ankit Siva and Samarth Wahal.

Lastly, I would like to thank my family for encouraging me to pursue my interests in math and computer science and giving me plenty of opportunities to do so.

Abstract

Distance preservers are a fundamental primitive used to sketch graphs and finite metric spaces. Despite having a variety of applications, bounds on the size of distance preservers for weighted, directed, acyclic graphs remain unimproved for almost 15 years [Coppersmith and Elkin *SIAM J. Disc. Math* '06, Szemeredi and Trotter *Combinatorica* '83]. We describe a novel approach to improve bounds on distance preservers in this setting using *path reroutings*. This formulation of the problem circumvents a known technical lower bound [Bodwin and Williams *SODA* '16]. Moreover, we consider a number of closely related problems and give evidence that the upper bounds from [Coppersmith and Elkin *SIAM J. Disc. Math* '06] are not tight.

Contents

1 Introduction

In this thesis, we describe a new approach to improve bounds on *distance preservers*.

Definition 1 (Distance Preservers)**.** *Given a weighted, possibly directed graph $G = (V, E, w)$ and a set of pairs $P \subseteq V \times V$, a distance preserver is a subgraph H of G such that $\text{dist}_H(v_1, v_2) = \text{dist}_G(v_1, v_2)$ for all $(v_1, v_2) \in P$.*

Distance preservers and other similar primitives in graph sketching such as distance oracles and spanners have found many applications in distributed computing and the all-pairs shortest paths problem [14, 17, 22]. A major open question asks how many edges are needed in a distance preserver on p pairs in a graph with n vertices. This question and other similar problems have been the object of study in a number of papers [1, 6, 8, 9, 10, 13].

In this thesis, we are primarily interested in improving upper bounds for distance preservers of weighted, directed, acyclic graphs. Before describing our new approach, we define some notation as developed in [7]:

Definition 2 (Path System)**.** *A path system is a pair (V, Π) where V is a finite ground set and Π is a non-empty set of sequences over V.*

Definition 3 (Strongly Metrizable)**.** *A path system (V, Π) is strongly metrizable if there exists a weighted, directed graph $G = (V, E, w)$ such that each $\pi \in \Pi$ is the unique shortest path between its endpoints in G.*

Definition 4 (Acyclic Path System)**.** *We say a path system (V, Π) is acyclic if there is an ordering of the ground set V such that every $\pi \in \Pi$ forms an increasing sequence.*

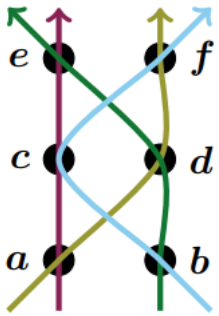

Figure 1: A acyclic path system that is not strongly metrizable with 4 paths: $\pi_b = bcf, \pi_g = bde, \pi_y = adf, \pi_r = ace$ and a ground set $V = \{a, b, c, d, e, f\}$. (Figure from [7]).

Note that to bound the size of a distance preserver with p demand pairs and n vertices, it suffices to bound the number of edges in a strongly metrizable path system with p paths and a ground set of size n. This is as after a small, random perturbation of the weights, we can assume that the shortest path between any pair of points is unique.

In past work, the primary property used about strongly metrizable shortest paths was *consistency*. Intuitively, consistency refers to the fact that the intersection of any two paths is a contiguous sub-path. Formally, we can define this as

Definition 5 (Consistency)**.** *A path system (V, Π) is consistent if for any two paths $\pi_1, \pi_2 \in \Pi$ and any two vertices $u, v \in V$ where u precedes v on both π_1 and π_2, the subpaths from u to v along π_1 and π_2 are equal.*

Using consistency, Coppersmith and Elkin showed in their original paper [13], which introduced distance preservers:

Theorem 1 (Coppersmith and Elkin [13]). *Let $G = (V, E, w)$ be a weighted, possibly directed graph on n vertices and $P \subseteq \binom{V}{2}$ a set of p demand pairs. Then G has a distance preserver with at most $O(n\sqrt{p})$ edges. Moreover, if G is undirected, then it has a distance preserver with at most $O(n+p\sqrt{n})$ edges.*

These are currently the state-of-the-art upper bounds known for weighted, directed, acyclic graphs. Moreover, Bodwin and Williams showed an example of p consistent shortest paths with $\Omega(\min\{n + p\sqrt{n}, n\sqrt{p}\})$ edges [9], implying that this analysis is tight if we only use consistency. Despite this, the best-known lower bounds for the problem only show that a distance preserver must have $\Omega((np)^{2/3}+n)$ edges [13, 30]. Moreover, the above upper bounds are known to not be tight for distance preservers of unweighted, undirected graphs [9].

In order to get an improvement on the bounds in Theorem 1, we will need to rely on structure beyond consistency. In [7], a complete characterization of strongly metrizable path systems is provided. However, it is quite technical and topological. Moreover, it's unclear how such a characterization may be used to get bounds. As such we consider *path reroutings*, which are more powerful than consistency and friendlier than the full characterization of strongly metrizable path systems.

Definition 6 (Path Rerouting). *Given an acyclic, edge-disjoint path systems (V, Π) and (V, Π') with paths $\pi_1, ..., \pi_p$ and $\pi'_1, ...\pi'_p$, we say that (V, Π') is a rerouting of (V, Π) if (1) for all i, π_i and π'_i start and end at the same vertices and (2) if u immediately proceeds v in some path $\pi'_i \in \Pi'$ then u immediately proceeds v in some path $\pi_j \in \Pi$. We call the rerouting trivial if $\pi_i = \pi'_i$ for all i.*

Given a path system (V, Π), we can create such a rerouting by allowing ourselves to swap paths at a given vertex in such a way that the rerouted path and original path start and end at the same vertices. As an example, in Figure 1, we can reroute the path π_r by swapping with π_y at a. We then follow π_y to d where we then swap with π_g. This gives us a rerouted path $\pi'_r = ade$. Now to reroute π_y we note that we already swapped with π_r at a; afterward, we can swap with π_b at c to get a rerouted path $\pi'_y = acf$. Continuing in this way, we get a rerouting $\pi'_r = ade, \pi'_b = bdf, \pi'_y = acf, \pi'_g = bce$.

The key fact that we will use is that a strongly metrizable path system does not have any non-trivial reroutings [7]. This gives us structure beyond consistency which we believe will yield improved bounds for distance preservers in the special case of directed, acyclic, edge-disjoint paths.

Formally, we will prove these bounds by considering *identity-free set systems*.

Definition 7 (Identity-Free Set System). *Given sets $A_1, A_2, ..., A_k \subseteq [n]$, we can define corresponding subgroups of the symmetric group of permutations over $[n]$, $H_1, ..., H_k$, where H_i is the set of all permutations that fix elements in $[n] \setminus A_i$. We then say that $A_1, ..., A_k$ is identity-free if for $h_i \in H_i$ and e denoting the identity permutation we have that $h_1 \circ h_2 \circ ... \circ h_k = e$, where $\circ$ denotes the composition of permutations, only has the trivial solution of $h_i = e$ for all i.*

We will be interested in bounding k, the number of sets, in terms of n, the size of the ground set, and s, the size of each subset, where we are assuming that $|A_i| = s$ for all i. Through the correspondence with path rerouting describe above, we have the following theorem:

Theorem 2. *Suppose (V, Π) is a strongly metrizable acyclic path system with edge-disjoint paths. Then there exists an identity-free set system with $|V|$ sets and $|\Pi|$ elements in the ground set such that each set has size $s = \Theta(|E|/|V|)$, where $|E|$ is the number of edges in a minimal (with respect to the number of edges) weighted graph realizing the path system.*

Combining the above theorem with well known acyclic path system lower bounds for distance preservers given in [13, 30] we have that:

Theorem 3. *For any positive integers n and s with $s \leq n$, there exists an identity-free set system $A_1, ..., A_k \subset [n]$ where each set has size $\Theta(s)$ and $k = \Omega\left(\frac{n^2}{s^3} + \frac{n}{s}\right)$.*

To get an upper bound, we note that since clearly no two distinct sets A_i and A_j can share a pair of elements, double counting pairs tells us that $k = O\left(n^2/s^2\right)$. This bound corresponds to using consistency in a path system and already gives an upper bound matching the best-known for the problem. However, as described previously, path reroutings and identity-free set systems are able to prove Figure 1 is not strongly-metrizable, even though it is consistent. In algebraic terms, we can write the identity as a product of distinct transpositions, which is not considered in this bound. As such, we believe that the upper bound of $O(n^2/s^2)$ is not tight.

Through Theorem 2, any bound for identity-free set systems of the form $k = O(n^2/s^c)$ when $s = O(\sqrt{n})$ implies a bound of $O(\min\{n^{1-1/c}p^{2/c} + n, n\sqrt{p}\})$ for distance preservers. Notably, for anything better than the folklore bound, i.e. for $c > 2$, we would get improvements over the bounds of Coppersmith and Elkin. Moreover, an upper bound matching that of Theorem 3 for identity-free set systems, would imply optimal bounds of $O((np)^{2/3} + n)$ for distance preservers in the special case of acyclic, edge-disjoint paths.

To further support that the bounds of Coppersmith and Elkin are not tight, we consider the following stability variant of identity-free set systems:

Definition 8 (Stable Identity-Free Set System)**.** *Given sets $A_1, ..., A_k \subset [n]$ we say they are a stable identity-free set system if for any permutation $\sigma \in S_k$ we have that $A_{\sigma(1)}, ...A_{\sigma(k)}$ is also an identity-free set system.*

Since this is a strictly stronger condition than being identity-free, we again have the consistency bound of $k = O(n^2/s^2)$. However, we show that this is polynomially far from tight by using the fact that the identity can be written as a product of distinct transpositions:

Theorem 4. *Suppose $A_1, ..., A_k \subseteq [n]$ are a stable identity-free set systems such that $|A_i| = s = O(\sqrt{n})$ for all i, then $k = O(n^2/s^{2.5})$.*

We also consider a variety of other closely related problems about identity-free groups and subsets. Specifically, we define these as

Definition 9 (Identity-Free Subsets)**.** *Given subsets $A_1, ..., A_k$ of a group G with identity element $e \in \bigcap_i A_i$ we say they are identity-free subsets if for $a_i \in A_i$ we have that $a_1a_2...a_k = e$ only has the trivial solution $a_i = e$ for all i.*

Definition 10 (Identity-Free Groups)**.** *Let G be a group with identity element e and $H_1, H_2, ..., H_k \leq G$ be subgroups of size s. We say $H_1, ...H_k$ forms an identity-free group system if for $h_i \in H_i$ the equation $h_1h_2...h_k = e$ has no non-trivial solutions.*

We believe that for identity-free subsets, the following conjecture holds:

Conjecture 1. *If $A_1, A_2, ..., A_k$ are identity-free subsets in a group G, where $|G| = n$ and $|A_i| = s$, then $k = O(n/s)$.*

Such a result would clearly be tight as given integers n and s we can take $G = \mathbb{Z}/n\mathbb{Z}$ and $A_i = \{0, 1, ..., s-1\}$ for all $1 \leq i \leq n/s$. Moreover, we make progress toward the conjecture by showing that $k = O((n/s)^c)$ for some $c < 2.622$ and that we indeed have $k = O(n/s)$ if G is abelian.

Now, for identity-free groups, one might expect significantly better bounds. Indeed if G is abelian, and $H_1, ..., H_k$ are identity-free groups then we have that $k = O(\log_s(n))$. In fact, one might hope that the key structure needed to prove bounds for identity-free set systems is that they correspond to identity-free groups. Clearly, in order for such an idea to work we would at the very least need to have that $k = O(\log(n)^c)$ for $s = \omega(1)$ and constant c. However, we show that this is not the case:

Theorem 5. *For infinitely many n, there exists a group G with $|G| = n$ and identity-free subgroups $H_1, ..., H_k \leq G$ where $|H_i| = s$, $k = 2^{\Omega(\sqrt{\log n})}$ and $s = 2^{\Omega(\sqrt{\log n})}$.*

In fact, we believe that even these bounds are far from optimal, and could see our constructions being generalized to show the existence of identity-free groups of size as large as $k = \Omega(n/s^2)$.

2 Preliminaries

2.1 Abstract Algebra

Since identity-free sets, groups, and subsets are fundamentally about the structure of groups, we recommend that the reader have a basic understanding of fundamental concepts in abstract algebra. A thorough treatment of these topics can be found in [18].

One definition that in particular will be quite important is the product set.

Definition 11. *(Product Set) If $H_1, H_2 \subseteq G$ then we define the product set of H_1 and H_2 as $H_1H_2 = \{h_1h_2 : h_1 \in H_1, h_2 \in H_2\}$.*

A key result about the product set tells that its size is large when H_1 and H_2 are subgroups of G, denoted by $H_1, H_2 \leq G$, and $H_1 \cap H_2$ is small.

Lemma 1. *If $H_1, H_2 \leq G$ then $|H_1H_2| = \frac{|H_1||H_2|}{|H_1 \cap H_2|}$.*

The proof is relatively straightforward and we refer an interested reader to Proposition 13 of [18].

2.2 The Fourier Transform on Finite Groups

One powerful, new tool that can now be used to prove bounds on distance preservers is representation theory. In this section, we briefly recall the basics of the Fourier transform on groups and give an example of how it can be used to prove bounds on the size of a product set. Since this is only a very high-level overview we refer the reader to [16, 27] for a more thorough treatment of the subject.

Given a finite-dimensional vector space V over $\mathbb{C}$ we will denote the group under multiplication of invertible matrices from V to V by $GL(V)$. We then define a representation of a group as

Definition 12 (Representation of a Group)**.** *A representation of a group G denoted by ρ is a homomorphism from G to $GL(V)$.*

The *dimension* of a representation which we denote by d_ρ is simply dimension of V. We call a representation *irreducible* if there is no subspace $0 \subsetneq W \subsetneq V$ such that $\rho(g)W \subseteq W$ for all $g \in G$. Finally, we call two representations $\rho : G \to GL(V_1)$ and $\tau : G \to GL(V_2)$ equivalent if they are the same after relabeling the elements in the vector space i.e. there is an injective linear map S from V_1 to V_2 such that $\tau(g)S = S\rho(g)$ for all $g \in G$.

Now given a representation, we can always split it into irreducible ones. That is

Theorem 6. *Let $\rho : G \to GL(V)$ be a representation. Then $V = \oplus_{i=1}^{m} W_i$ where the W_i's are orthogonal subspaces, $\rho(g)W_i \subseteq W_i$ for all $g \in G$, and $\rho|^{W_i} : G \to GL(W_i)$ are irreducible representations.*

We can also assume $\rho(g)$ is always unitary. Finally, we have that any finite group G only has finitely many irreducible representations and more precisely

$$\sum_{\rho} d_\rho^2 = |G|$$

where the sum is taken over all irreducible representations.

As an example of an irreducible representation, every group has a *trivial representation*: Let V be a one-dimensional vector space over $\mathbb{C}$ and define ρ as $\rho(g)x = x$ for all $x \in V$ and $g \in G$.

Now given $f : G \to \mathbb{C}$ we can define it's Fourier transform, which maps a representation to a linear transformation.

$$\hat{f}(\rho) = \sum_{g \in G} f(g)\rho(g)$$

Such a transformation is useful as it works well with convolution, which is extended to groups in the natural way:

$$f_1 * f_2(s) = \sum_{a,b \in G : ab = s} f_1(a) f_2(b)$$

With the above definition of the Fourier transform one can check that

$$\widehat{f_1 * f_2}(\rho) = \hat{f}_1(\rho)\hat{f}_2(\rho)$$

We also get other extentions of classical Fourier tools such as Pareseval's identity and Fourier inversion. Before we define them we define the Frobenius inner product for square matrices of the same size as $\langle X, Y \rangle = \mathrm{tr}(XY^*)$ where Y^* is the conjugate transpose of Y. We denote $\|X\|^2 = \|X\|_F^2 = \langle X, X \rangle$. We now have *Parseval's identity*:

$$\sum_{g \in G} |f(g)|^2 = \frac{1}{|G|} \sum_{\rho} d_\rho \|\hat{f}(\rho)\|^2$$

And the *Fourier inversion formula*:

$$f(g) = \frac{1}{|G|} \sum_{\rho} d_\rho \langle \hat{f}(\rho), \rho(g) \rangle$$

One strength of Fourier methods for groups is that they naturally capture the problem of finding the size of a product set. To illustrate this we show the following result of Gowers:

Theorem 7 (Gowers [11])**.** *Let $m(G)$ be the minimum dimension of any non-trivial irreducible representation. If $|A||B||C| > |G|^3/m(G)$ then $ABC = G$.*

Proof. The proof follows simply using Fourier methods and Cauchy Schwartz. We consider the indicator function $1_A, 1_B, 1_C$ for each set and define $f = 1_A * 1_B * 1_C$. We will show that $f(g) > 0$ for all $g \in G$, which clearly implies that $ABC = G$. Applying Parseval's identity to 1_A gives

$$|A| = \frac{1}{|G|} \sum_{\rho} d_\rho \|\widehat{1_A}(\rho)\|^2$$

So for any non-trivial representation ρ we can throw away all the remaining terms to get that

$$\|\widehat{1_A}(\rho)\| \leq \sqrt{\frac{|A||G|}{m(G)}}$$

Denote the trivial representation as 1. Fourier inversion applied to f gives us that

$$\begin{aligned}
f(g) = \frac{1}{|G|} \sum_{\rho} d_\rho \langle \hat{f}(\rho), \rho(g) \rangle &\geq \frac{|A||B||C|}{|G|} - \frac{1}{|G|} \sum_{\rho \neq 1} d_\rho \langle \hat{f}(\rho), \rho(g) \rangle \\
&\geq \frac{|A||B||C|}{|G|} - \frac{1}{|G|} \sum_{\rho \neq 1} d_\rho \|1_A(\rho)\| \cdot \|1_B(\rho)\| \cdot \|1_C(\rho)\| \\
&\geq \frac{|A||B||C|}{|G|} - \sqrt{\frac{|A||G|}{m(G)}} \sqrt{\frac{1}{|G|} \sum_{\rho \neq 1} d_\rho \|1_B(\rho)\|^2} \cdot \sqrt{\frac{1}{|G|} \sum_{\rho \neq 1} d_\rho \|1_C(\rho)\|^2} \\
&\geq \frac{|A||B||C|}{|G|} - \sqrt{\frac{|A||G|}{m(G)}} \sqrt{|B|} \sqrt{|C|}
\end{aligned}$$

Which is positive since $|A||B||C| > |G|^3/m(G)$. □

Since identity-free sets are fundamentally a question about products of subgroups, we have that this powerful, new toolbox at our disposal to extract information from the problem. We describe how this might be done more closely in Section 3.3.

2.3 Path System Reroutings

Finally, for completeness, we include a proof that strongly metrizable path systems have no non-trivial reroutings.

Lemma 2 (Bodwin [7])**.** *Let (V, Π) be an acyclic, edge-disjoint path system. If (V, Π) has a non-trivial rerouting then it is not strongly metrizable.*

Proof. Let (V, Π) and (V, Π') be acyclic, edge-disjoint path systems with paths $\pi_1, ..., \pi_p$ and $\pi'_1, ..., \pi'_p$ respectively, where (V, Π') is a non-trivial rerouting of (V, Π). Towards a contradiction, suppose (V, Π) is strongly metrizable. It then follows that there is a graph $G = (V, E, w)$ such that each $\pi \in \Pi$ is the unique shortest path between its endpoints. Now denote by E_1 the set of edges in G used by some path $\pi \in \Pi$. Similarly, denote by E_2 the set of edges in G used by some path $\pi' \in \Pi'$. By the second condition of being a rerouting we have that $E_2 \subseteq E_1$ and thus

$$\sum_{e \in E_1} w(e) \geq \sum_{e \in E_2} w(e)$$

But now note that by the first condition of being a rerouting and the fact that π_i are shortest path in G

$$\sum_{e \in \pi_i} w(e) \leq \sum_{e \in \pi'_i} w(e)$$

Moreover, since this is a non-trivial rerouting there exists an i such that $\pi_i \neq \pi'_i$. But now since π_i is the unique shortest paths in G between its endpoints

$$\sum_{e \in \pi_i} w(e) < \sum_{e \in \pi'_i} w(e)$$

Using that the paths are edge-disjoint then implies

$$\sum_{e \in E_1} w(e) = \sum_{i=1}^{p} \sum_{e \in \pi_i} w(e) < \sum_{i=1}^{p} \sum_{e \in \pi'_i} w(e) = \sum_{e \in E_2} w(e)$$

a contradiction. □

3 Identity-Free Set Systems

3.1 Connections To Path Systems

In this section, we prove many of results stated in the introduction. Before doing so, however, we show that we can assume that all sets in an identity-free set system have size exactly s:

Lemma 3. *If there exists an identity-free set system $A_1, ..., A_k \subset [n]$ with $\frac{1}{k}\sum_{i=1}^{k} |A_i| = s$ then there exists an identity-free set system $B_1, ..., B_{k'}$ where every set B_i has size exactly $s' = \Omega(s)$ and $k' = \Omega(k)$.*

Proof. We note that for any set A_i with $|A_i| \geq 2s$ we can write $A_i = C_1 \cup C_2$ where C_1 and C_2 are disjoint, $|C_1| = s$, and $|C_2| \geq s$. Moreover the new set system $A_1, A_2, ..., A_{i-1}, C_1, C_2, A_{i+1}, ...A_k$ is also identity-free.

We can continue splitting large sets in this way. Let $k_{\geq s}$ and $k_{<s}$ denote the number of set of size at least s and less than s respectively. If we ever have that $k_{\geq s} \geq k/2$ then we can simply take our B_i's to be the sets of size at least s. We can clearly remove elements to make every set have size exactly s.

We can alternatively stop once there are no longer any sets of size at least $2s$. In this case, we split at most $k/2$ sets since $k_{\geq s}$ increases each time we split a set and we ended with $k_{\geq s} < k/2$. Thus,

there are at most $3k/2 \leq 2k$ sets. We then note that the average number of elements in a set is at least

$$\frac{\sum_i |A_i|}{2k} \geq s/2$$

Now we will remove any set with fewer than $s/4$ elements. Let k' be the number of sets remaining. Then we have that

$$\frac{2s \cdot k' + (2k - k') \cdot s/4}{2k} \geq \frac{\sum_i |A_i|}{2k} \geq s/2$$

$$k' \geq 2/7k$$

Again, we can delete elements from the sets so that we have that every set has size exactly $\lfloor s/4 \rfloor$. Taking these as our B_i's proves the lemma. □

Leveraging the above result, we can now prove the connection between path systems and permutations described in the introduction.

Theorem 2. *Suppose (V, Π) is a strongly metrizable acyclic path system with edge-disjoint paths. Then there exists an identity-free set system with $|V|$ sets and $|\Pi|$ elements in the ground set such that each set has size $s = \Theta(|E|/|V|)$, where $|E|$ is the number of edges in a minimal (with respect to the number of edges) weighted graph realizing the path system.*

Proof. We will denote $|\Pi| = p$ and $|V| = n$. Since the path system is acyclic and strongly metrizable, there is an ordering of the vertices such that all paths are strictly increasing sequences. We will enumerate the vertices in this order $v_1, v_2, ..., v_n$. We also fix some arbitrary labeling of the paths $\pi_1, ..., \pi_p$.

Now define $A_i = \{j : v_i \in \pi_j\}$. We claim that such a set system is identity-free. To see this, suppose not. Note that if $A_1, ..., A_n$ is not identity-free then the set system obtained by reversing the order of the sets i.e. $A_n, A_{n-1}, ..., A_1$ is also not identity-free. So let $\sigma_n ... \sigma_1$ be a non-trivial way of writing the identity. Now for each path π_i we will create a rerouted path π_i'. Initially, π_i' will start at the same point as π_i. We then proceed to build π_i' as follows: If we are at v_j then the next term in π_i' will be term after v_j in $\pi_{\sigma_j ... \sigma_1(i)}$ assuming v_j isn't the final term in that path. If it is the final term in $\pi_{\sigma_j ... \sigma_1(i)}$, then we stop.

For this to we well-defined we first need to show the following claim:

Claim: $\pi_{\sigma_j ... \sigma_1(i)}$ goes through v_j
Proof. We proceed by induction. Let v_a be the first term of the sequence π_i. Then it follows that $\sigma_a \sigma_{a-1} ... \sigma_1(i) = \sigma_a(i)$ since we have that $i \notin A_k$ for $k < a$. Hence, since A_a only contains paths through a, the statement holds. Now let the mth element in the rerouted sequence π_i' be v_t and let $v_{t'}$ be the vertex immediately before it in π_i'. It follows that $\sigma_{t-1} ... \sigma_1(i) = \sigma_{t'} ... \sigma_1(i)$. Hence, if σ_t doesn't move $\sigma_{t-1} ... \sigma_1(i)$ then the path $\pi_{\sigma_t \sigma_{t-1} ... \sigma_1(i)}$ contains v_t since $\pi_{\sigma_{t'} ... \sigma_1(i)}$ did. On the other hand, if σ_t does move $\sigma_{t-1} ... \sigma_1(i)$ then it must be sent to a path in A_t which by definition contains v_t. □

So, the rerouting process is in fact well-defined. Next, we show

Claim: π_i' ends at the same vertex as π_i
Proof. Denote the last vertex in π_i by v_f. Towards a contradiction, suppose we stopped on some other vertex v_t. Then we must have ended while trying to follow a path π_j with $j \neq i$ that ended. But note that since π_j has ended, we have that $j \notin A_k$ for $k > t$. But then $\sigma_n \sigma_{n-1} ... \sigma_1(i) = j \neq i$, which is a contradiction. □

Applying this transformation to every path in (V, Π) we get a rerouting (V, Π'), which we claim has edge-disjoint paths.

Claim: (V, Π') has edge-disjoint paths

Proof. Suppose not, and assume that π'_i and π'_j share a pair of consecutive vertices (v_a, v_b). Since (V, Π) is edge-disjoint we have that (v_a, v_b) is a consecutive pair of vertices in exactly one path π_k. But now note that this implies that $\sigma_a\sigma_{a-1}...\sigma_1(i) = k$ and $\sigma_a\sigma_{a-1}...\sigma_1(j) = k$, which is a contradiction. □

Since the representation of the identity is non-trivial, we have found a non-trivial rerouting of (V, Π) but this contradicts Lemma 2. Thus the sets A_i are identity free. Using Lemma 3 then gives us the desired result.

□

Using the correspondence between path systems and identity-free set systems, along with a lower bound to the Szemerédi-Trotter Theorem we can prove Theorem 3 from the introduction.

Theorem 3. *For any positive integers n and s with $s \leq n$, there exists an identity-free set system $A_1, ..., A_k \subset [n]$ where each set has size $\Theta(s)$ and $k = \Omega\left(\frac{n^2}{s^3} + \frac{n}{s}\right)$.*

Proof. The proof follows by a slight modification to the proof given in [21]. For simplicity, assume $4s^2|n$. Consider a grid of points in the plane $[n/(4s^2)] \times [n/s]$. Then consider all lines of the form $ax + b$ where $a \in [2s]$ and $b \in [n/(2s)]$. Then we can easily see that each line intersects exactly $n/(4s^2)$ points. Hence we have that there are $n^2/(4s^3)$ points, n lines, and an average of s incidences per point.

There is now a clear correspondence with an acyclic edge-disjoint unique shortest path system, where the points are the ground sets and the lines are the shortest paths. So using Theorem 2 we have that the above corresponds to a set system with $n^2/(4s^3)$ sets and a ground set of size n where each set is of size $\Theta(s)$.

Note we can trivially find identity-free set systems with $k = \Omega(n/s)$ by simply taking our sets to be disjoint sets of size s. □

3.2 Upper Bounds for Identity-Free Set Systems

We now proceed to show that we can easily prove bounds matching the best-known from Theorem 1.

Lemma 4 (Folklore Upper Bound)**.** *For any identity-free set system $A_1, ..., A_k \subseteq [n]$ with $|A_i| = s \geq 2$ for all i, we have that $k = O\left(\frac{n^2}{s^2}\right)$.*

Proof. Note that an identity-free set system must satisfy $|A_i \cap A_j| \leq 1$ for $i \neq j$. By double counting pairs of elements appearing in a set, we then have

$$k\binom{s}{2} = \sum_{i=1}^{k} \binom{s}{2} \leq \binom{n}{2}$$

which implies that $k = O\left(\frac{n^2}{s^2}\right)$. □

If s is large we can also give an optimal bound that matches the lower bound given in Theorem 3.

Lemma 5. *Let $A_1, ..., A_k \subseteq [n]$ be an identity-free set system where $|A_i| = s$ and suppose $s \geq 2e\sqrt{n}$, then $k = O\left(\frac{n}{s}\right)$.*

Proof. Define $B_i = A_i \setminus (\bigcup_{j=1}^{i-1} A_j)$. Now since $B_i \subseteq A_i$, $B_1, ..., B_k$ also forms an identity-free set system. Let $R_1, ..., R_k$ be the corresponding groups for the sets $B_1, ..., B_k$. Since $B_1, ...B_k$ are disjoint it follows that the elements in $R_1...R_k$ commute and hence that $R_1, R_1R_2, R_1R_2R_3, ...$ are all groups. Moreover, since they are identity-free we have that $R_i \cap R_1...R_{i-1} = \{e\}$.

Suppose now that $k > s$. Then we have that by Lemma 1

$$|R_1R_2...R_k| \geq s!(s-1)!...1! \geq ((s/2)!)^{s/2} \geq \left(\frac{s}{2e}\right)^{s^2/4}$$

On the other hand we have that this product must be at most $n!$ so we have that

$$\left(\frac{s}{2e}\right)^{s^2/4} \leq n! \leq n^n$$
$$(s^2/8)\log(n) \leq n\log(n)$$

Taking n sufficiently large then clearly gives us that $s < 3\sqrt{n}$, which is a contradiction.

So, we must have that $k < s$. Considering the size of $R_1R_2...R_k$ then gives

$$s!(s-1)!...(s-k)! \leq n^n$$
$$\left(\frac{s}{2e}\right)^{sk/4} \leq n^n$$

Taking the log of both sides then clearly gives us that $k = O(n/s)$ as desired. □

Together, these allow us to prove bounds matching those given in [13].

Lemma 6. *Let $G = (V, E, w)$ be a directed, acyclic graph and P be a set of pairs of vertices of G. If the shortest paths between pairs of vertices in P are edge-disjoint, then we have that there exists a distance preserver with at most $O(\min\{|V| + |P|\sqrt{|V|}, |V|\sqrt{|P|}\})$ edges.*

Proof. Note that the paths in the preserver correspond to an acyclic, edge-disjoint path system with $|P|$ paths and a ground set of size $|V|$. Let $|E_H|$ be the number of edges in a minimal preserver.

By Theorem 2, this corresponds to an identity-free set system with $|V|$ sets and a ground set of size $|P|$. Moreover, each set will have size $\Theta(|E_H|/|V|)$. Now if $|E_H|/|V| = \Omega(1)$ then we can apply Lemma 4. This then gives us that

$$|V| = O\left(\frac{|P|^2|V|^2}{|E_H|^2}\right)$$
$$|E_H| = O\left(|P|\sqrt{|V|}\right)$$

Hence it follows that since we assumed that $|E_H| = \Omega(|V|)$ we must have that

$$|E_H| = O(|P|\sqrt{|V|} + |V|)$$

Now to get the second bound we note that if $|E_H|/|V| = O(\sqrt{|P|})$ then we trivially have that $|E_H| = O(|V|\sqrt{|P|})$. So we assume that $|E_H|/|V| = \Omega(\sqrt{|P|})$. But then Lemma 5, tells us $|V| = O(|V||P|/|E_H|)$. So, $|E_H| = O(|P|) = O(|V|\sqrt{|P|})$ as desired. □

We note that improving Lemma 4 immediately yields improved bounds via an analogous argument. So, the regime of interest will be $s = O(\sqrt{n})$.

3.3 Representation Theory

Finally, we discuss the formulation of the problem in terms of Fourier analysis. Since we are simply outlining how these tools might be useful to prove bounds on distance preservers we omit proofs and only provide sketches.

A fundamental result that we will need to compute Fourier transforms is

Lemma 7 (Schur's Lemma)**.** *Let $\rho : G \to GL(V_1)$ and $\tau : G \to GL(V_2)$ be representations of G and S be a linear map from V_1 to V_2 such that for all $g \in G$*

$$\tau(g)S = S\rho(g)$$

Then if ρ and τ are not equivalent $S = 0$. Additionally, if $V_1 = V_2$ and $\rho = \tau$ then S is a constant multiple of the identity.

This immediately gives us the following result

Lemma 8. *Let G be a finite group and U be the uniform distribution over G i.e. $U(g) = 1/|G|$ for all $g \in G$. Then for any irreducible representation $\rho : G \to GL(V)$*

$$\hat{U}(\rho) = \begin{cases} 1 & \rho \text{ is trivial} \\ 0 & \text{otherwise} \end{cases}$$

Sketch of Proof. Note that $\hat{U}(\rho) = 1/|G| \sum_{g \in G} \rho(g)$. Now for any $h \in G$ we have that $\rho(h^{-1})\hat{U}(\rho)\rho(h) = \hat{U}(\rho)$. The result then follows by Schur's lemma and the observation that $\rho(g)\hat{U}(\rho) = \hat{U}(\rho)$ for all $g \in G$. □

Lemma 9. *Let G be a finite group and U_H denote the uniform distribution over a subgroup H i.e. $U_H(h) = 1/|H|$ for $h \in H$ and 0 otherwise. Then for any representation $\rho : G \to GL(V)$ we have that $\hat{U}(\rho)$ is an orthogonal projection.*

Sketch of Proof. We consider the representation of H induced by ρ, $\rho|_H : H \to GL(V)$. We then decompose V into orthogonal invariant subspaces $W_1, ..., W_m$ under $\rho|_H$. Using Lemma 8, it then follows that $\hat{U}$ is either the 0 map on W_i if the character restricted to W_i, $\rho|_H^{W_i}$ is not trivial, or 1 if it is trivial. This immediately implies $\hat{U}(\rho)$ is an orthogonal projection onto the space

$$\bigoplus_{i : \rho|_H^{W_i} \text{ is trivial}} W_i$$

□

We now describe how one might extract information about identity-free sets using representation theory. Let $A_1, ..., A_k \subset [n]$ be an identity-free set system where each set has size s and let $H_1, ..., H_k$ be the corresponding groups. We note that the subgroups corresponding to an identity-free set system are conjugate. Let g_i be defined such that $H_i = g_i H_1 g_i^{-1}$. Thus if U_i denotes the uniform distribution over H_i we have that

$$\hat{U}_i(\rho) = \rho(g)\hat{U}_1\rho(g^{-1})$$

So all the matrices are unitary equivalent. Using this we can now encode our problem into the language of Fourier analysis. Specifically, we can write the identity-free condition as:

Theorem 8. *Let $A_1, ...A_k \subseteq [n]$ and H_i be the corresponding groups that fix elements from $[n] \setminus A_i$. If U_1 denotes the uniform distribution over H_1 and g_i are such that $H_i = g_i H_1 g_i^{-1}$ then $A_1, ..., A_k$ is identity-free if and only if*

$$\frac{1}{(n!)^k} = \frac{1}{n!} \sum_{\rho} d_\rho \operatorname{tr}\left(\prod_i \rho(g_i)\hat{U}_1\rho(g_i^{-1}) \right)$$

This already naturally captures various properties of identity-free sets. For instance, since the trace is invariant under cyclic permutations we have that the same holds for the ordering of identity-free sets.

Of course, by Theorem 5, this is not enough to prove the desired bounds for identity-free set systems, although it may give improved bounds for identity-free groups. However, we note that the representations of both the symmetric group and the specific subgroups in question are known [24]. Thus, it is quite feasible for additional information to be extracted.

Additionally, it seems to be the case that transpositions are quite important for giving good bounds for identity-free set systems. A non-trivial estimate on the number of transpositions in $H_1...H_k$ would likely yield improved bounds. In this formulation, counting this quantity is quite natural. Specifically, we use the following corollary of Schur's lemma from [16]

Lemma 10 (Diaconis [16]). *Let f be a function that is constant on conjugacy classes and $\rho : G \to GL(V)$. Then $\hat{f}(\rho) = \lambda_\rho I$, where $I \in GL(V)$ is the identity map.*

Applying this to the indicator function on transpositions then gives us that the number of transpositions in $H_1 H_2 ... H_k$ is

$$|G|^{k-1} \sum_\rho d_\rho \lambda_\rho \operatorname{tr}\left(\prod_i \rho(g_i) \hat{U}_1 \rho(g_i^{-1})\right)$$

where we again have that the values of λ_ρ are explicitly known [16]. This has a striking resemblance to the identity-free constraint.

4 Stable Identity-Free Sets

As mentioned previously, the success of our approach for distance preservers relies on the fact that we can write the identity as a product of distinct transpositions and that this will lead to asymptotic improvements on our upper bounds. In this section, we show that we can in fact get significant improvements on a closely related problem precisely because of a non-trivial way of writing the identity as a product of transpositions.

We consider a variant of the identity-free problem where we allow ourselves to shuffle the sets.

Definition 13 (Stable Identity-Free Set System). *Given sets $A_1, ..., A_k$ we say they are a stable identity-free set system if for any permutation $\sigma \in S_k$ we have that $A_{\sigma(1)}, ...A_{\sigma(k)}$ is identity-free.*

We are again interested in giving a non-trivial bound on k in terms of s and n. Clearly, this is a strictly stronger condition than that of identity-free permutations. Thus, we have the same bounds as before of $k = O(n^2/s^2)$ and $k = O(n/s)$ if $s = \Omega(\sqrt{n})$. However, we show that a stronger bound holds.

Theorem 9. *Suppose $A_1, ..., A_k$ are a stable identity-free set-systems and $s = O(\sqrt{n})$, then $k = O(n^2/s^{2.5})$.*

This strongly indicates that this approach can yield improved bounds in the ordered setting as well. The key structure we use in our proof is that we can write the identity as a non-trivial product of transpositions i.e. $(12)(34)(13)(24)(14)(32) = e$. This corresponds precisely to the set system arising from Figure 1.

To prove the theorem we define the transposition graph:

Definition 14 (Transposition Graph). *Let $A_1, ..., A_k$ be an unordered identity-free set system. The corresponding transposition graph is an edge colored graph on the vertex set $[n]$. There is an edge between i and j of color m if $\{i, j\} \subseteq A_m$.*

By consistency, since we have that $A_1, ..., A_k$ are identity-free it follows that every edge has exactly one color. We will bound the number of four cycles in such a graph, which immediately implies a bound on k through standard cycle saturation bounds. While we choose to use the transposition graph, the bounds in this section can also be attained by considering the bipartite incidence graph. This is a graph with vertices corresponding to elements of $[n]$ and the sets A_i with an edge between an element a and a set A_i if $a \in A_i$.

To prove our bounds, we will need to work with a more well-behaved graph. So recall the following useful lemma:

Lemma 11. *Given a graph $G = (V, E)$ with average degree d there exists an induced subgraph with minimum degree $d/4$.*

Proof. Start out with $S = V$. If there is a vertex with degree smaller than $d/4$ in $G[S]$ remove it from S. Note that we removed strictly less than $nd/4 = |E|/2$ edges. So this process terminates with a non-empty subset S and $G[S]$ gives the desired graph. □

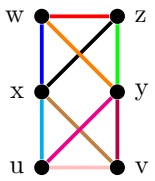

Figure 2: A Rainbow Door

We note that since the graph is induced, we have that the regularized transposition graph also corresponds to a stable, identity-free set system $B_1, ..., B_{k'} \subset [n']$, where we removed empty sets and elements appearing in no set B_i. Clearly, we have that $k' = \Omega(k)$; moreover, we will assume that $n' = \Omega(n)$, as having many isolated vertices allows us to give strictly better bounds. So we now assume that our identity-free set systems are *normalized*:

Definition 15 (Normalized Identity-Free Set System)**.** *An identity-free set system $A_1, ..., A_k$ is normalized if each set has size at most s and the minimum degree in the transposition graph is $\Omega(ks^2/n)$.*

We now show the following simple lemma:

Lemma 12. *Let $A_1, ..., A_k$ be a normalized, identity-free set system and G its transposition graph. If S is an independent set in G then $|S| = O(n/s)$.*

Proof. Let $S = \{s_1, ..., s_m\}$ be an independent set in G and define $T_i = \{j : s_i \in A_j\}$. Since S is an independent set, no set contains both s_i and s_j for $i \neq j$, so $T_i \cap T_j = \emptyset$. Now using the minimum degree in G and that a set contains at most s elements, we have $|T_i| = \Omega(ks/n)$. But then

$$\Omega(mks/n) = \sum_{i=1}^{m} |T_i| \leq k$$

which immediately implies $m = O(n/s)$. □

The key structure we will use is that the transposition graph cannot contain a *rainbow door* in the transposition graph (See Figure 2).

Lemma 13. *If $A_1, ..., A_k$ are a stable identity-free set system and G is the corresponding transposition graph then G does not contain a rainbow door.*

Proof. For the sake of contradiction, suppose G contained a rainbow door as depicted in Figure 2. Then note that

$$(xw)(yz)(wy)(xz)(wz)(ux)(vy)(vx)(uy)(uv) = e$$

which contradicts our assumption that the A_i's form a stable identity-free set system. □

With this, we can now bound the number of four cycles in the transposition graph. The rough idea is as follows: Consider the common neighbors of vertices x and y. Since the transposition graph doesn't contain any rainbow doors, there must not be too many edges between these common neighbors. We then use our bound on the size of an independent set to show that the number of common neighbors of x and y must be small, which implies a bound on the number of four cycles.

Lemma 14. *Let $A_1, ..., A_k$ be a stable, normalized, identity-free set system and G its transposition graph. Moreover assume that $s = O(\sqrt{n})$ and $k = \Omega(n/s)$, then there are at most $O(n^4/s^2)$ four cycles in G.*

Proof. Consider two vertices x and y. We will show that they have at most $O(n/s)$ common neighbors, which immediately implies the claim.

First, suppose there is a set A_i that contains both x and y and denote it by $A_{x,y}$. If no such set exists let $A_{x,y} = \emptyset$. Clearly, in either case $|A_{x,y}| \leq s = O(n/s)$.

Now let S denote the common neighbors of x and y that are not in $A_{x,y}$. We'll proceed to bound the size S by arguing that it's close to an independent set. We begin by removing any set A_i containing $\{v, x\}$ or $\{v, y\}$ for some $v \in S$. This results in a new stable identity-free set system $B_1, .., B_\ell$ with transposition graph H. Note that this process removes at most 2 sets per vertex, as if $u \in A_i$ and $\{v, x\}$ caused the removal of A_i then $\{u, x\} \subseteq A_i$. Hence, we have that H has minimum degree $\Omega(ks^2/n - 2s) = \Omega(ks^2/n)$.

Now if S forms an independent set, then by Lemma 12 we have that $|S| = O(n/s)$ and we've controlled the size of S. So suppose there are vertices $w, z \in S$ such that $wz \in E(H)$. Then since $w, z \notin A_{x,y}$ and we didn't remove the edge wz when removing sets, there is no set A_i containing three elements of $\{x, y, w, z\}$. This implies that the edges wx, wy, wz, zx, zy all have distinct colors. Now, denote by $A_1, ..., A_5$ the sets containing $\{w, x\}, \{w, y\}, \{w, z\}, \{z, x\}, \{z, y\}$.

Consider the vertices in $R = S \setminus \cup_{i=1}^{5} A_i$. Towards a contradiction, suppose there is an edge between $u, v \in R$. As before, we must have that ux, uy, uv, vx, vy have distinct colors. Moreover, since we removed the vertices incident to edges of the same colors as wx, wy, wz, zx, zy we have that the colors of $ux, uy, uv, vx, vy, wx, wy, wz, zx, zy$ are all distinct. But then u, v, w, x, y, z form a rainbow door, which is a contradiction. Hence, R is an independent set in H and $|S| \leq 5s + |R| = O(5s + n/s) = O(n/s)$ by Lemma 12.

Since all the common neighbors of x and y are contained in $A_{x,y} \cup S$ which we showed has size $O(n/s)$, we have proven the claim. □

We now recall the following cycle saturation bound

Lemma 15 (Folklore)**.** *Let $G = (V, E)$ be a graph with average degree d. If $d = \Omega(\sqrt{n})$ with a large enough implicit constant then G contains $\Omega(d^4)$ four cycles.*

Proof. Let $x_{u,v}$ denote the number of paths of length two from u to v and c_4 the number of four cycles. We then clearly have that

$$c_4 = \sum_{\{u,v\} \in \binom{V}{2}} \binom{x_{u,v}}{2} = \sum_{\{u,v\} \in \binom{V}{2}} \Omega(x_{u,v}^2 - 1) = \sum_{\{u,v\} \in \binom{V}{2}} \Omega(x_{u,v}^2) - O(n^2)$$

Using Cauchy-Schwartz then gives

$$\begin{aligned}
&\geq \Omega\left(\frac{\left(\sum_{\{u,v\} \in \binom{V}{2}} x_{u,v}\right)^2}{n^2}\right) - O(n^2) \\
&= \Omega\left(\frac{\left(\sum_{v \in V} \binom{\deg(v)}{2}\right)^2}{n^2}\right) - O(n^2) \\
&= \Omega\left(\frac{\left(\sum_{v \in V} \deg(v)^2\right)^2}{n^2}\right) - O(n^2) \\
&= \Omega\left(\frac{E^4/n^2}{n^2}\right) - O(n^2) = \Omega(d^4) - O(n^2)
\end{aligned}$$

where the last line follows by another application of Cauchy-Schwartz. □

We can combine the above with Lemma 14 to prove the theorem:

Proof of Theorem 9. By Lemma 11 we can assume the set-system is regular and has minimum degree $\Omega(ks^2/n)$. Now note that if $ks^2 = O(n^{3/2})$ then $k = O(n^{3/2}/s^2) = O(n^2/s^3)$, in which case we are

done. So, we assume we have $\Omega(n^{3/2})$ edges as required by Lemma 15. We then have that there are at least

$$\Omega((ks^2/n)^4) = \Omega(k^4s^8/n^4)$$

four cycles. Comparing to Lemma 14 we have that

$$\Omega(k^4s^8/n^4) = O(n^4/s^2)$$

Giving us that $k = O(n^2/s^{2.5})$ as desired. □

We end this section by remarking that our bounds for stable identity-free sets apply to a more local version of the problem: Let $A_1, ...A_k \subset [n]$ with $|A_i| = s$ be a set system (not necessarily identity-free) with the property that any subsystem of 10 sets from $A_1, ..., A_k$ form a stable identity-free set system. Then $k = O(n^2/s^{2.5})$. As such, using larger forbidden structures or more generally using a more global property could yield better bounds.

5 Identity-Free Subsets

We now turn to prove bounds on a more general problem of identity-free subsets. These will immediately imply upper bounds on identity-free groups. Recall

Definition 9 (Identity-Free Subsets)**.** *Given subsets $A_1, ..., A_k$ of a group G with identity element $e \in \bigcap_i A_i$ we say they are identity-free subsets if for $a_i \in A_i$ we have that $a_1a_2...a_k = e$ only has the trivial solution $a_i = e$ for all i.*

As with the other variants of identity-free problems, we will assume that all of the sets have size s and attempt to bound the size of k in terms of n, which denotes the size of the group G, and s.

We begin by noting that we indeed have that k must be finite:

Lemma 16 (Trivial Bound)**.** *For identity-free sets, $A_1, ..., A_k \subseteq G$ where $|A_i| = s$ and $|G| = n$, we have that $k = O(n^2/s)$.*

Proof. Note that we have that any element $g \in G$ can appear at most n times as $g^n = e$. It follows that

$$ks \leq n^2$$

giving the bound. □

It's easy to see the we can in fact have that every non-identity element has order n, for instance in $\mathbb{Z}/p\mathbb{Z}$ for p prime. However, we'll show that this bound is not tight for s large. Specifically, we'll show that $k = O((n/s)^c)$ for some $c > 1$. We note that we can have $k = \Omega(n/s)$ as taking $G = \mathbb{Z}/n\mathbb{Z}$ and $A_1 = A_2 = ... = A_k = \{0, 1, ..., s-1\}$ with $k = n/s$ gives us identity-free subsets. We conjecture that this is indeed tight

Conjecture 1. *If $A_1, A_2, ..., A_k$ are identity-free subsets in a group G, where $|G| = n$ and $|A_i| = s$, then $k = O(n/s)$.*

This seems to be the correct bound and implies that the above simple bound is never tight.

5.1 Ruzsa Distance

In order to prove our bounds for general groups, we will utilize a useful distance from additive combinatorics, which was shown naturally extends to groups by Tao [31].

Definition 16 (Ruzsa Distance)**.** *Given a group G and $A, B \subseteq G$ we will let*

$$d(A, B) = \log\left(\frac{|AB^{-1}|}{|A|^{1/2}|B|^{1/2}}\right)$$

As expected, the Ruzsa distance has a number of properties that are morally required of any good distance function. Specifically, it is non-negative, symmetric, and translation invariant, and satisfies the triangle inequality.

Lemma 17 (Tao [31])**.** *The Ruzsa distance is non-negative, symmetric, translation invariant, and satisfies the triangle inequality. That is if G is a group, $A, B, C \subseteq G$, and $x \in G$ then $d(A,B) \geq 0$, $d(A,B) = d(B,A)$, $d(x \cdot A, x \cdot B) = d(A,B)$, and $d(A,C) \leq d(A,B) + d(B,C)$.*

Note, however, that it does not satisfy $d(A,A) = 0$ for all $A \subseteq G$.

Proof. It is easy to see that $d(x \cdot A, x \cdot B) = d(A,B)$ holds. Additionally, clearly $|AB^{-1}| \geq \max(|A|,|B|) \geq \sqrt{|A||B|}$, giving positivity. Symmetry follows from the fact that $(AB^{-1})^{-1} = BA^{-1}$, so $|AB^{-1}| = |BA^{-1}|$.

So now all that remains is the triangle inequality. It suffices to show

$$|AC^{-1}||B| \leq |AB^{-1}||BC^{-1}|$$

We prove this by defining an injection $f : AC^{-1} \times B \to AB^{-1} \times BC^{-1}$. Given an $x \in AC^{-1}$ there may be many $a \in A, c \in C$ such that $x = ac^{-1}$. For each element in $x \in AC^{-1}$ fix one such decomposition. Now given an $(x,b) \in AC^{-1} \times B$ we map it to (ab^{-1}, bc^{-1}), where $x = ac^{-1}$ is the fixed decomposition of x. To see it's injective suppose $f(x,b_1) = f(w,b_2)$ where x has fixed decomposition $a_1c_1^{-1}$ and w has fixed decomposition $a_2c_2^{-1}$. Then note that,

$$x = (a_1b_1^{-1})(b_1c_1^{-1}) = (a_2b_2^{-1})(b_2c_2^{-1}) = w$$

This implies that $a_1 = a_2$ and $c_1 = c_2$ which in turn implies $b_1 = b_2$, giving us that f is an injection. Thus the desired inequality follows. □

5.2 Improved Upper Bounds

Using the Ruzsa distance, we can now show that the product of identity-free subsets is large, which will give us a non-trivial bound on the size of identity-free subsets.

Lemma 18. *Let $A, B \subseteq G$ with $e \in A \cap B$ and suppose that when $a \in A$ and $b \in B$, the equation $ab = e$ only has the solution $a = b = e$. If $|A|, |B| \geq t$ then we have that $|AB| \geq ct - 1$ where $c = \frac{\sqrt{13}}{2} - \frac{1}{2} \approx 1.30$*

Proof. Suppose not. Then we have that

$$|AB| < ct \leq c|A|^{1/2}|B|^{1/2}$$

Hence it follows that

$$d(A, B^{-1}) = \log\left(\frac{|AB|}{|A|^{1/2}|B|^{1/2}}\right) \leq \log(c)$$

Thus we have that

$$d(A,A) \leq d(A,B^{-1}) + d(B^{-1},A) = 2d(A,B^{-1}) \leq 2\log(c)$$

Hence

$$d(A,A) = \log\left(\frac{|AA^{-1}|}{|A|}\right) \leq \log(c^2)$$
$$|AA^{-1}| \leq c^2|A|$$

Now since $e \in A \cap A^{-1}$ we have that $A \cup A^{-1} \subseteq AA^{-1}$.

$$|A| + |A^{-1}| - |A \cap A^{-1}| = |A \cup A^{-1}| \leq |AA^{-1}| \leq c^2|A|$$
$$|A \cap A^{-1}| \geq (2 - c^2)|A|$$

Now note that we must have that $(A \cap A^{-1}) \cap B = \{e\}$. Clearly if there was any other element g in the intersection then $g^{-1}g = e$ where $g^{-1} \in A$ and $g \in B$, which is a contradiction. Since $A \cup B \subseteq AB$

$$|AB| \geq |A \cap A^{-1}| + |B| - 1 \geq (3 - c^2)t - 1$$

But now note that this is a contradiction since we chose c such that $3 - c^2 = c$. □

This immediately implies that

Theorem 10. *If $A_1, ..., A_k \subseteq G$ are identity-free subsets with $|G| = n$ and $|A_i| = s$ then $k = O((n/s)^c)$ where $c = \log(2)/\log(\sqrt{13}/2 - 1/2) \approx 2.621$.*

Proof. Let $N(m)$ denote the number of elements in the product $A_1A_2...A_m$. By Lemma 18 we have that

$$N(m) \geq (\sqrt{13}/2 - 1/2)N(m/2) - 1$$

which implies that

$$N(m) = \Omega\left(sm^{1/c}\right)$$

where $c = \log(2)/\log(\sqrt{13}/2 - 1/2)$. Now clearly $N(k) \leq n$ so

$$\Omega(sk^{1/c}) \leq n$$
$$k = O((n/s)^c)$$

as desired. □

We briefly note that we can get better bounds for non-abelian finite simple groups using Theorem 7 since the degree of the minimum non-trivial representation is $\omega(1)$ [11]. We can also get an improved bound if $A_1 = A_2 = ... = A_k$ by strengthening Lemma 18.

Lemma 19. *Let $A \subseteq G$ with $e \in A$ and suppose that for $a_1, a_2 \in A$ the equation $a_1a_2 = e$ only has the solution $a_1 = a_2 = e$. Then $|AA| \geq \sqrt{2}|A| - 1$.*

Proof. Suppose not and assume $|AA| < (\sqrt{2} - 1/|A|)|A|$. Following the same lines as that of Lemma 18 we have that

$$|A \cap A^{-1}| \geq (2 - (\sqrt{2} - 1/|A|)^2)|A|$$
$$|A \cap A^{-1}| \geq 2\sqrt{2} - 1/|A| > 1$$

But on the other hand, note that we clearly must have that $A \cap A^{-1} = \{e\}$. So, we reached a contradiction. □

This gives us that

Theorem 11. *If $A_1, ..., A_k \subseteq G$ are identity-free subsets of size s with $|G| = n$ and $A_1 = A_2 = ... = A_k$ then $k = O((n/s)^2)$.*

While the bounds in this section can likely be improved, it seems to be the case that $O((n/s)^2)$ is a natural barrier due to the loss incurred by using the triangle inequality. One might hope that if $|AA^{-1}| < (2 + \epsilon)|A|$ then we still have that $|A \cap A^{-1}|$ is large. However, we remark that this is not the case. To see this simply take a subgroup H with no elements of order two and partition it into two sets A, A^{-1} such that $H = A \cup A^{-1}$ and $A \cap A^{-1} = \{e\}$. Then $|AA^{-1}| = |H| \leq 2|A|$. Regardless, bounds of the form $k = O(n^2/s^2)$ would still be interesting as they imply that the trivial bound can only be tight for constant s.

5.3 Optimal Bounds for Abelian Groups

Further supporting the fact that Conjecture 1 is true, we prove that it holds when G is abelian.

Theorem 12. *Suppose $A_1, ..., A_k \subseteq G$ are identity-free subsets of size s with $|G| = n$, then if G is abelian $k = O(n/s)$.*

We show this by improving Lemma 18 in the abelian setting. Key in the proof of this result is Kneser's Theorem. Before we state the theorem, we first define

Definition 17 (Stabilizer)**.** *The stabilizer of a set A is the set $S(A) = \{g \in G : A + g = A\}$.*

It is easy to verify that $S(A)$ is a group.

Theorem 13 (Kneser's Theorem)**.** *Suppose $A, B \subseteq G$ are finite, non-empty sets and G is an abelian group. Moreover, let $H = S(A + B)$. Then $|A + B| \geq |A + H| + |B + H| - |H|$*

There are relatively short proofs known [15], but we do not include them here. This immediately implies the improved version of Lemma 18 and gives us a proof of Theorem 12.

Corollary 1. *Let $A, B \subseteq G$ with G abelian and $0 \in A \cap B$, Moreover assume that for $a \in A$ and $b \in B$ the equation $a + b = 0$ only has the solution $a = b = 0$. If $|A|, |B| \geq t$ then $|A + B| \geq 2t - 1$.*

Proof. By Kneser's Theorem, we have that $|A + B| \geq |A + H| + |B + H| - |H|$. Since $e \in A \cap B$ we have that $H \cup A \subseteq A + H$ and $H \cup B \subseteq B + H$. We'll show that

$$|A \cup H| + |B \cup H| \geq |A| + |B| + |H| - 1$$

It clearly suffices to show that

$$|H \setminus B| + |H \setminus A| \geq |H| - 1$$

Let $g \in H \cap A$ where $g \neq 0$ then we must have $g^{-1} \notin H \cap B$. Hence $(H \cap A)^{-1} \setminus \{0\} \subseteq H \setminus B$ and it follows that $|H \setminus B| \geq |H \cap A| - 1$, which immediately gives us the desired conclusion. □

6 Identity-Free Groups

Finally, we prove bounds on the size of identity-free groups.

Definition 10 (Identity-Free Groups)**.** *Let G be a group with identity element e and $H_1, H_2, ..., H_k \leq G$ be subgroups of size s. We say $H_1, ...H_k$ forms an identity-free group system if for $h_i \in H_i$ the equation $h_1 h_2 ... h_k = e$ has no non-trivial solutions.*

Again we assume that $|A_i| = s$ for all i and $|G| = n$. Our main result in this section is that k is not polylogarthimically bounded. Specifically

Theorem 5. *For infinitely many n, there exists a group G with $|G| = n$ and identity-free subgroups $H_1, ..., H_k \leq G$ where $|H_i| = s$, $k = 2^{\Omega(\sqrt{\log n})}$ and $s = 2^{\Omega(\sqrt{\log n})}$.*

Perhaps most importantly, this implies that the specific structure of groups in our problem on identity-free set systems is vital to proving the desired bounds.

For an upper bound the number of identity-free groups we can have, we combine Lemma 1 with Theorem 10 to get that $k = O((n/s^2)^c)$ for some $c < 2.622$. Assuming Conjecture 1, this can be improved to $k = O(n/s^2)$. Despite differing quite significantly from our lower bounds, the bound of $k = O(n/s^2)$ may very well be tight. In fact our construction for Theorem 5 relies on showing that for $s = 2$ we can in fact find $\Omega(n/s^2)$ identity-free groups. In some sense, this implies the product of subgroups in the non-abelian setting and arbitrary subsets containing the identity can behave similarly.

6.1 Lower Bounds

We will construct a set of identity-free groups in the dihedral group D_{2n}. Recall that this group is defined by two generators r, s which satisfy $r^n = e$, $s^2 = e$, and $srs = s^{-1}$. We can generalize this last relation to see that $sr^i s = srs(sr^{i-1}s) = r^{-i}$.

Lemma 20. *There exist groups $H_1, ..., H_n \leq D_{2n}$ with $|H_i| = 2$ such that $h_1 h_2 ... h_n = e$ only has the trivial solution for $h_i \in H_i$.*

Proof. We will take $H_i = \langle r^i s \rangle$, the cyclic subgroup generated by $r^i s$. Since $r^i s r^i s = e$, $|H_i| = 2$ as desired. Now suppose the groups are not identity-free and let $i_1 < i_2 < ... < i_j$ be the subindices such that $h_{i_l} \neq e$ in some non-trivial representation of the identity i.e.

$$h_{i_1} h_{i_2} ... h_{i_j} = e$$
$$r^{i_1} s r^{i_2} s ... r^{i_j} s = e$$

If j is odd then

$$r^{i_1 - i_2 + i_3 ... - i_{j-1} + i_j} s \neq e$$

So assume j is even. Then we must have that

$$i_1 - i_2 + i_3 - i_4 + ... i_{j-1} - i_j = 0 \mod n$$

Since the sequence $i_1, i_2, ..., i_j$ is increasing we observe

$$(i_1 - i_2) + (i_3 - i_4) + ... (i_{j-1} - i_j) < 0$$

On the other hand,

$$i_1 + (-i_2 + i_3) + (-i_4 + i_5) ... + (-i_{j-2} + i_{j-1}) - i_j \geq i_1 - i_j \geq 1 - n > -n$$

But then we cannot have that $i_1 - i_2 + i_3 - i_4 + ... + i_{j-1} - i_j = 0 \mod n$ and have reached a contradiction. □

We can then use the following black box result to extend the above bound to give a lower bound for larger s. However, we incur some cost increasing s, and likely don't get optimal results.

Lemma 21. *Let m be a positive integer and $H_1, ..., H_k$ be identity-free groups in G where $H_i = s$ and $|G| = n$. Then we can find identity-free groups $H'_1, ..., H'_k$ in a group G' where $|H'_i| = s^m$ and $|G'| = n^m$*

Proof. We will work in the group $G^m = G \times G \times ... \times G$. Define $H'_i = H_i^m = H \times H \times ... \times H$. Now suppose

$$h'_1 ... h'_k = e$$

is a non-trivial representation of the identity. Denote $h'_i = (h_i^{(1)}, ..., h_i^{(m)})$ where each $h_i^{(j)} \in H_i$. Now choose some $h'_i \neq e$. This means that we just have that $h_i^{(j)} \neq e$ for some j. But then

$$h_1^{(j)} h_2^{(j)} ... h_k^{(j)} = e$$

is a non-trivial representation of the identity. But this contradicts the fact that the H_i's are identity-free groups. □

Theorem 5. *For infinitely many n, there exists a group G with $|G| = n$ and identity-free subgroups $H_1, ..., H_k \leq G$ where $|H_i| = s$, $k = 2^{\Omega(\sqrt{\log n})}$ and $s = 2^{\Omega(\sqrt{\log n})}$.*

Proof. Applying Lemma 21 to our dihedral group lower bound, for any positive integers a, b we get identity-free group systems with $k = a$, groups of size 2^b in a group of size $|G| = (2a)^b$. Taking $b = \log(a)$, we get $k = a$ groups of size a with $|G| = (2a)^{\log(a)}$ giving the desired bound. □

7 Conclusion

In this thesis, we considered a promising approach to bound the size of distance preservers using identity-free set systems. We presented various bounds drawing from results in abstract algebra, extremal combinatorics, and additive combinatorics. However, many of the bounds we presented are not tight or at the very least do not have matching lower bounds leaving several unanswered questions.

Perhaps, the most important unresolved question is:

Question 1. *If $A_1, ..., A_k \subseteq [n]$ form an identity-free set system with $|A_i| = s$, then does $k = O(n^2/s^3 + n/s)$? Moreover, if we ask that $A_1, ..., A_k$ are stable identity-free sets do tighter bounds hold?*

Additionally, we ask

Question 2. *Does Conjecture 1 hold? That is for identity-free subsets $A_1, ..., A_k \subseteq G$ with $|A_i| = s$ and $|G| = n$ do we always have that $k = O(n/s)$?*

Finally, we ask whether it's indeed the case that the general bound for identity-free subgroups is only marginally better than that of identity-free subsets.

Question 3. *For all positive integers s, n with $s < \sqrt{n}$, does there exist a group G and identity-free groups $H_1, H_2, ..., H_k \leq G$, where $|G| = n$, $|H_i| = s$, and $k = \Omega(n/s^2)$? If not, what are the correct bounds for the problem?*

References

[1] Abboud, A., and Bodwin, G. Error amplification for pairwise spanner lower bounds. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms* (2016), SIAM, pp. 841–854.

[2] Abboud, A., and Bodwin, G. The 4/3 additive spanner exponent is tight. *Journal of the ACM (JACM) 64*, 4 (2017), 28.

[3] Alon, N. Testing subgraphs in large graphs. *Random Structures & Algorithms 21*, 3-4 (2002), 359–370.

[4] Alon, N., and Shapira, A. Testing subgraphs in directed graphs. *Journal of Computer and System Sciences 69*, 3 (2004), 354–382.

[5] Baswana, S., Kavitha, T., Mehlhorn, K., and Pettie, S. Additive spanners and (α, β)-spanners. *ACM Transactions on Algorithms (TALG) 7*, 1 (2010), 5.

[6] Bodwin, G. Linear size distance preservers. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms* (2017), Society for Industrial and Applied Mathematics, pp. 600–615.

[7] Bodwin, G. On the structure of unique shortest paths in graphs. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms* (2019), Society for Industrial and Applied Mathematics, pp. 2071–2089.

[8] Bodwin, G., Grandoni, F., Parter, M., and Williams, V. V. Preserving distances in very faulty graphs. *arXiv preprint arXiv:1703.10293* (2017).

[9] Bodwin, G., and Williams, V. V. Better distance preservers and additive spanners. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms* (2016), Society for Industrial and Applied Mathematics, pp. 855–872.

[10] Bollobás, B., Coppersmith, D., and Elkin, M. Sparse distance preservers and additive spanners. *SIAM Journal on Discrete Mathematics 19*, 4 (2005), 1029–1055.

[11] Breuillard, E. A brief introduction to approximate groups. *Thin groups and superstrong approximation 61* (2014), 23–50.

[12] Chandra, A. K., Furst, M. L., and Lipton, R. J. Multi-party protocols. In *Proceedings of the fifteenth annual ACM symposium on Theory of computing* (1983), ACM, pp. 94–99.

[13] Coppersmith, D., and Elkin, M. Sparse sourcewise and pairwise distance preservers. *SIAM Journal on Discrete Mathematics 20*, 2 (2006), 463–501.

[14] Cowen, L. J. Compact routing with minimum stretch. *Journal of Algorithms 38*, 1 (2001), 170–183.

[15] DeVos, M. A short proof of kneser's addition theorem for abelian groups. In *Combinatorial and Additive Number Theory*. Springer, 2014, pp. 39–41.

[16] Diaconis, P. Group representations in probability and statistics. *Lecture notes-monograph series 11* (1988), i–192.

[17] Dor, D., Halperin, S., and Zwick, U. All-pairs almost shortest paths. *SIAM Journal on Computing 29*, 5 (2000), 1740–1759.

[18] Dummit, D. S., and Foote, R. M. *Abstract algebra*, vol. 3. Wiley Hoboken, 2004.

[19] Elkin, M. An improved construction of progression-free sets. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms* (2010), SIAM, pp. 886–905.

[20] Gowers, W. T. A new proof of szemerédi's theorem. *Geometric and Functional Analysis 11*, 3 (2001), 465–588.

[21] Jukna, S. *Extremal combinatorics: with applications in computer science*. Springer Science & Business Media, 2011.

[22] Peleg, D., and Upfal, E. A trade-off between space and efficiency for routing tables. *Journal of the ACM (JACM) 36*, 3 (1989), 510–530.

[23] Pyber, L. How abelian is a finite group? In *The Mathematics of Paul Erdös I*. Springer, 1997, pp. 372–384.

[24] Sagan, B. E. *The symmetric group: representations, combinatorial algorithms, and symmetric functions*, vol. 203. Springer Science & Business Media, 2013.

[25] Sanders, T. Solving xz= y 2 in certain subsets of finite groups. *The Quarterly Journal of Mathematics 68*, 1 (2017), 243–273.

[26] Serra, O., Vena, L., et al. A combinatorial proof of the removal lemma for groups. *Journal of Combinatorial Theory, Series A 116*, 4 (2009), 971–978.

[27] Serre, J.-P. *Linear representations of finite groups*, vol. 42. Springer, 1977.

[28] Solymosi, J. Roth-type theorems in finite groups. *European Journal of Combinatorics 34*, 8 (2013), 1454–1458.

[29] Solymosi, J. The (7, 4)-conjecture in finite groups. *Combinatorics, Probability and Computing 24*, 4 (2015), 680–686.

[30] Szemerédi, E., and Trotter, W. T. Extremal problems in discrete geometry. *Combinatorica 3*, 3-4 (1983), 381–392.

[31] Tao, T. Product set estimates for non-commutative groups. *Combinatorica 28*, 5 (2008), 547–594.